Navy Personnel Research and Development Center
San Diego, California 92152-7250 TN-96-21 February 1996

The Estimation of Incremental Validity in an Employment Setting:

A Review of Relevant Literature and Some Research Design Considerations

Norman G. Peterson
Mary Ann Hanson
John H. Wolfe

19960325 128

Approved for public release; distribution is unlimited.

NPRDC-TN-96-21 February 1996

The Estimation of Incremental Validity in an Employment Setting: A Review of Relevant Literature and Some Research Design Considerations

Norman G. Peterson
Mary Ann Hanson
Personnel Decisions Research Institute

John H. Wolfe
Navy Personnel Research and Development Center

Reviewed by
Gerald J. Laabs

Approved and released by
Kathleen E. Moreno
Director, Personnel and Organizational Assessment

Approved for public release;
distribution is unlimited.

Navy Personnel Research and Development Center
53335 Ryne Road
San Diego, CA 92152-7250

Blank Pages

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	February 1996	Final—September 1988-May 1989

4. TITLE AND SUBTITLE
The Estimation of Incremental Validity in an Employment Setting: A Review of Relevant Literature and Some Research Design Considerations

5. FUNDING NUMBERS
Program Element: O&M,A Reimbursable
Work Unit: MIPR 88-T-104
Contract: N66001-87-D-0085

6. AUTHOR(S)
Norman G. Peterson, Mary Ann Hanson, & John H. Wolfe

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Personnel Decisions Research Institute
43 Main Street, SE, Suite 405
Minneapolis, MN 55414

8. PERFORMING ORGANIZATION REPORT NUMBER
NPRDC-TN-96-21

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Navy Personnel Research and Development Center
53335 Ryne Road
San Diego, CA 92152-7250

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Functional Area: Personnel
Product Line: Printed Testing
Product Line: ASVAB

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Improved validity coefficients have been reported for retest ASVAB scores. Two hypotheses have been made to explain this effect: Temporal Contiguity (TC), which posits a change in true scores over time, and Better Ability Estimate (BAE), which posits a reduction in error of measurement, but no true score change. This report critically reviews literature relevant to the evaluation of these two hypotheses and proposes a research design to investigate the hypotheses. The evaluation of the viability of either hypothesis is viewed as partly dependent on the type and number of criteria chosen to be predicted, and the timing of data collection on the criteria. The probability of true score changes occurring is viewed as varying within and across predictor domains (cognitive, perceptual/psychomotor, biodata, personality, vocational interests). The probability of BAE explaining validity increases is viewed as partly dependent on the nature of interventions to reduce error of measurement. A research design to investigate the two hypotheses directly is described and power analyses are provided for the design. Finally, it is suggested that more than one study will be necessary to provide firm conclusions about the TC versus BAE hypothesis, and it might be more prudent to design incremental validity research to avoid confounding with these issues.

14. SUBJECT TERMS
Military selection, employment test validation, ASVAB, incremental validity, decay of validity, dynamic criteria

15. NUMBER OF PAGES
52

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

Foreword

This report addresses a controversial issue in the design of experiments to determine the improvement in validity that can be obtained from adding new aptitude tests to the Armed Services Vocational Aptitude Battery (ASVAB). The issue is whether the ASVAB needs to be readministered at the same time as the new tests are administered. This issue is of practical importance because of the costs and difficulties of increased testing time in research.

This effort was conducted under the Future Testing Project sponsored by the Office of the Assistant Secretary of Defense (Force Management & Personnel, Military Manpower & Personnel Policy) with U. S. Army Operations and Maintenance funds (O&M,A Reimbursable; Work Unit MIPR 88-T-104). The report was written under Contract N66001-87-D-0085, Task Order 7J19 with Personnel Decisions Research Institute. John J. Pass was the Contracting Officer's Technical Representative (COTR) for the contract, and John H. Wolfe was the Assistant COTR for the task.

Lloyd Humphreys and Rodney Rosse provided invaluable advice and assistance on this effort. Dr. Humphreys provided insight and counsel based on his many years of innovative research in the area of individual differences, prediction, and change. Dr. Rosse was instrumental in clarifying the research design and model suggested for investigating the primary hypotheses. Thanks also to John Novak, who provided his usual professional, efficient effort in the production of the manuscript.

KATHLEEN E. MORENO
Director, Personnel and Organizational Assessment

Summary

Problem

Research directed toward improving the prediction of successful performance of Navy enlisted personnel requires the estimation of additions from experimental predictor tests to the validity achieved by the Armed Services Vocational Aptitude Battery (ASVAB). Even very small additions to validity can provide large increases in utility or productivity of the selected personnel. This estimation of additional, or incremental validity, is complicated by the fact that experimental tests cannot always be administered contemporaneously with the ASVAB, and raises the issue of whether or not the ASVAB should be readministered with the experimental tests.

Some research has shown that a second administration of the ASVAB provides scores with higher validity for predicting success than the first administration. Two technical explanations or hypotheses have been put forth for this finding. The first explanation, labeled Temporal Contiguity (TC), hypothesizes that persons' scores on the variables measured by the ASVAB change over time, and thus the ASVAB scores closest in time to the criterion to be predicted will show the highest validity. The second explanation, labeled Better Ability Estimate (BAE), hypothesizes that a second administration of the ASVAB removes some of the error in measuring persons' scores on the ASVAB variables, and thus results in a more accurate measure of the "true score" for the person—but the "true score" does not change. If the first of these hypotheses is true, then all predictors (ASVAB and experimental) should be administered at the same time in order to provide accurate estimates of incremental validity. If the second hypothesis is true, then predictors can be administered at different times, but should be administered the same number of times.

Objectives

This paper has two main objectives. The first is to review the available evidence relevant to the two competing hypotheses we have just described (TC versus BAE). This review will provide an initial estimate of the plausibility of the hypotheses for estimating incremental validities for experimental tests over the ASVAB when predicting performance of first-term Navy enlistees. The second objective is to recommend possible research designs and associated analyses that will shed further light on the plausibility of the two hypotheses, providing comments on avoiding the confounding of the TC and BAE hypotheses.

Approach

The literature was searched by using computerized databases, queries to active researchers, and manual reviews of recent journals likely to have relevant literature. The articles and reports were reviewed and appropriate conclusions were drawn with regard to the problem, especially with regard to useful research designs to investigate the hypotheses.

Results and Discussion

The review of the literature led to the conclusion that there is not a preponderance of evidence to support one hypothesis over the other, although the likelihood of the TC hypothesis being supported for the ASVAB in the Navy selection setting is not viewed as high. There is much

evidence that the validity of ability tests for predicting various criteria decays over time, though the credibility of the evidence is questioned by some researchers, and the meaning of this decay is not unanimously agreed upon. The type of test (cognitive, personality, etc.), the type of criteria, and the timing of data collection could all have an impact on the viability of the TC versus BAE hypothesis.

A research design is described that will allow a direct test of the TC and BAE hypotheses, with sufficient power to detect validity increments of .01. One research study will probably not provide enough firm information to conclude that TC or BAE does or does not explain any increases in ASVAB validities, primarily because only one type of criterion and one time interval between ASVAB administrations can be investigated. For this reason, it is suggested that it may be most prudent to design incremental validation research so that TC and BAE hypotheses cannot confound the interpretation of results.

Contents

List of Figures

Introduction

Incremental Validity and Timing of Predictor Administrations

Researchers investigating experimental tests of traits of job applicants are often interested in the increases in validity that occur over and above the validity obtained by the use of the current selection instruments. This increase in validity, which is called *incremental validity*, can be affected by the timing of the administration of the experimental tests relative to the operational tests. Ideally, all the candidate tests should be administered as close as possible in time. Pragmatically, this is often not possible. Operational tests are usually administered to all applicants, who are screened based on their scores on the tests. In almost all cases, only those passing the screens and choosing to join the organization are available to take the experimental tests, usually some time after the operational tests were administered. Unless the operational tests are readministered, performance on the experimental tests may be differentially influenced by whatever events have occurred between the two times of administration.

Stability of Predictor Scores and Predictor/Criterion Relationships

If scores on tests and instruments measuring psychological traits were perfectly stable, and the relationships between those scores and scores on important criteria (such as success in training, supervisory ratings of job performance, achievement on job knowledge or performance tests) were also perfectly stable, then differences in the timing of administration of tests would be unimportant. In this paper, we review evidence strongly indicating that such scores and relationships are not perfectly stable. For example, Humphreys (1986) notes:

> If a necessary characteristic of a psychological trait is being fixed at a stable level over long periods of time, there are probably no psychological traits. . . . The human organism is dynamic. Instability occurs during normal development without experimental intervention. The research questions that are of immediate concern are the degrees of stability manifested by the various traits that psychologists are interested in. (p. 4)

More specific to the interests of this paper, research by the Center for Naval Analyses (Mayberry, 1988) shows higher validities for retest scores for the Armed Services Vocational Aptitude Battery (ASVAB) than for initial ASVAB scores. The research question that needs to be addressed is: Why did the difference in validity for the original and retest ASVAB scores occur?

The Temporal Contiguity (TC) and Better Ability Estimate (BAE) Hypotheses

Two competing hypotheses have been advanced to explain the difference in validity for the original and retest ASVAB scores. The first hypothesis is that the retest scores have higher validity because persons' true scores change over time, and, since the retest scores are administered closer in time to the collection of criterion data, they will show a stronger relationship to the criterion. This hypothesis can be called the Temporal Contiguity (TC) hypothesis. The competing hypothesis states that the retest scores have higher validity because there is less error in the retest score, but persons' true scores have not changed from the first testing. This hypothesis can be called the Better Ability Estimate (BAE).

If the TC hypothesis is correct (or mostly correct), then studies that administer some measures later in time than others may lead to higher estimates of the absolute and incremental validity of the later-administered measures. If we assume that the operational measures are administered at the appropriate time for selection purposes and that this time is earlier than when the experimental tests are administered, then the operational tests will be at a disadvantage when evaluation of contributions to prediction of criteria is made. The implication for the research question is that the research design should include the concurrent administration (or as nearly concurrent as possible) of all candidate measures.

If the BAE hypothesis is correct (or mostly correct), then concurrent administration of candidate measures is unimportant. Instead, all candidate measures should be administered the same number of times so that equal opportunity for reduction in error will prevail. More pragmatically, perhaps, steps should be taken to identify and reduce the error present in the first administration of candidate measures.

If both hypotheses are correct to some degree, or differentially correct with regard to various kinds of candidate predictors, various research design and analyses may be required to adequately answer the research question.

In evaluating the plausibility of these two hypotheses, one factor to keep in mind is the timing of the criterion measurement. Concurrent administration of some of the candidate predictors with the criterion would result in shared measurement error and, therefore, higher validities than if the same predictors were administered at a different time than the criteria (whether or not they were administered concurrently with the other candidate predictors).

Objectives

This paper has two main objectives. The first is to review the available evidence relevant to the two competing hypotheses we have just described (TC versus BAE). This review will provide an initial estimate of the plausibility of the hypotheses for estimating incremental validities for experimental tests over the ASVAB when predicting performance of first-term Navy enlistees. The second objective is to recommend possible research designs and associated analyses that will shed further light on the plausibility of the two hypotheses, providing comments on avoiding the confounding of the TC and BAE hypotheses.

Approach

Literature Search

We attempted to locate relevant literature from civilian and military sources, and from published and unpublished sources. First, we performed a computerized search of the PsycInfo database, obtaining over a dozen sources that appeared relevant and reviewing these publications to generate additional sources. We also reviewed editions for the last five years of the following journals: *Psychological Bulletin, Journal of Applied Psychology, Personnel Psychology, Journal of Educational Psychology, Educational and Psychological Measurement, Applied Psychological Measurement,* and *Organizational Behavior and Human Decision Processes (formerly Organizational Behavior and Human Performance).* Finally, we queried researchers especially

knowledgeable in topics related to the research question on incremental validity for any references or publications, published or unpublished, that they thought would be useful.

Trait Stability and Intervening Events

The principal data reviewed here are correlations between test scores over time, in particular, patterns of correlations called "simplex" or "quasi-simplex." Humphreys (1960) discusses the properties of correlation matrices that have this simplex structure and their implications for understanding the stability of individual differences over time. Simplex matrices may occur when a single variable is measured at multiple points in time and these scores are intercorrelated. These correlation matrices fit the simplex structure when change from one point in time to another is unrelated to initial score. If this is true, the following relationship will hold:

$$\rho_{x_i x_k} = \rho_{x_i x_j} \rho_{x_j x_k}, \text{ where } i < j < k.$$

This is equivalent to saying that if a score obtained at a point in time between two other scores is partialled out of the correlation between the two separated scores, the resulting partial correlation will be zero. Actual test scores never fit this simplex structure perfectly because of measurement error, and correlation matrices based on actual test scores are sometimes referred to as quasi-simplex matrices. This means that scores that change quite a bit from one administration to the next will have lower test-retest reliabilities than scores that have changed very little. Note that this is not necessarily true if the intercorrelations of test scores do not fit the simplex pattern. Scores that do not fit the simplex pattern could change a great deal from one administration to the next and still have high stability coefficients. (For example, if the test scores increase a great deal overall, but for higher ability examinees the increase is greater than for lower ability examinees, the resulting stability coefficients could be quite high. The resulting pattern of intercorrelations, however, would not fit the simplex.) Since most longitudinal ability data seem to fit the simplex, we can generally expect scores to be more stable over shorter periods of time. Since change will be independent of initial scores, we can expect scores that are changing rapidly over time to have lower stability coefficients than those that are changing more slowly. A comparison of stability coefficients for children versus adults supports this model, since intelligence test scores are increasing fairly rapidly for children, and these scores have correspondingly lower stability coefficients than those of adolescents and adults.

The likelihood of true score changes (i.e., the TC hypothesis being correct) and/or reduction in error of measurement (i.e., the BAE hypothesis being correct) may interact with the type of predictor being considered. For example, most cognitive ability measures might be very stable in the population of interest (that is, young adults aged about 18 to 25), but personality scales might be relatively less stable. Thus, the TC hypothesis would be more plausible for explaining validity differences for personality measures, but the BAE hypothesis would be more plausible for cognitive measures. The military services are considering new predictors from several domains (Campbell, 1986; Wolfe, 1988). Therefore, we review relevant literature in the major domains of likely predictor measures—cognitive, perceptual, psychomotor, personality, biodata, and vocational interests. We also review the effects of different intervening events on test scores.

Decreases in Validity Over Time

Many studies that have directly investigated changes in validity with increases in time between predictor and criterion administrations show decreasing validity for predictor measures over time. There are some exceptions. We review the literature for this topic, discussing its implications for validation research.

Nature of the Criterion

Although the nature of the criterion is not mentioned in the two competing hypotheses for explaining the validity of predictors, the criterion variable(s) chosen for a study can affect the evaluation of the two hypotheses.

First, if criterion performance is changing over time, requiring different ability mixes at different points in the performance curve, then the point at which criterion measurement is made may greatly affect the observed relationships with candidate abilities. This is sometimes referred to as the "changing task" explanation for some observed ability-task performance relationships (Alvares & Hulin, 1973).

Second, the type of criterion task may affect the observed relationships with candidate abilities. Ackerman (1987) posits that certain task characteristics (e.g., "consistency") interact with the type of predictor measure and this affects the pattern of validities over time for each of the various predictors.

Job performance, in the broader sense, is often thought of as multidimensional (Campbell, 1986; Schmidt & Kaplan, 1971). If the several dimensions have different relationships with candidate predictors, and a single dimension is chosen (wittingly or otherwise) as the criterion, then, obviously, a set of limitations has been imposed on the predictor/criterion relationships that can be observed. For example, in the Army's Project A research, five dimensions of job performance were identified, and the observed incremental validities of various experimental predictors to the ASVAB ranged from .00 to .21 (Campbell, 1986, pp. 8-10).

We review some of the literature pertinent to these points, evaluating the implications of these criterion issues and identifying possible ways of addressing the issues in research design and analysis of data.

Validation Design and Statistical Issues

Design and statistical issues are discussed at several points in the paper. We deal with these issues as they arise in our review of the literature. Various validation research designs have been directly discussed (Sussmann & Robertson, 1986), and the strengths and weaknesses of these designs and their feasibility for Navy use are evaluated.

Finally, Wolfe (1988) has identified and discussed several issues associated with the estimation of incremental validity for experimental predictors over the ASVAB. We have found little to add to his discussion.

Trait Stability and Intervening Events

The Stability of Psychological Traits

The purpose of this section is to review and summarize available evidence concerning the extent to which true scores for cognitive abilities and other individual differences are likely to change. The TC hypothesis can only be correct if there actually are true score changes in the abilities of interest. The extent to which these true score changes can affect the validity of scores is limited by the size of the changes. Since the only type of change that will affect the validity of scores is a change in the rank order of scores, this review will focus on the test-retest reliability of scores where the two administrations are separated by a substantial period of time. These long term test-retest coefficients will be referred to as "stability coefficients." Where possible, the discussion will focus on young adults (ages 18 to 25), since most military applicants are likely to be in this age group. Stability of a trait could also be interpreted to mean no change in the mean level of scores for a group of examinees over time, but a change in mean level will have no direct effect on the validity of scores. A discussion of differences in mean scores typically found between different age groups will be presented in the next section, but in the context of the effect of intervening events, particularly maturation, on test scores.

Literature concerning the stability of cognitive abilities will be presented first. To the extent that the available studies of cognitive abilities deal with abilities similar to those measured by the ASVAB, this will provide an estimate of how much we could expect scores for the ASVAB to change over time. Literature will be presented on the stability of other types of individual differences as well, since the armed services are currently considering new predictors from several other domains. Additional categories of individual differences that will be discussed are perceptual and psychomotor abilities, vocational interests, personality traits, and biodata.

Where possible, the immediate or short term test-retest correlations for scores will be presented in order to put stability coefficients into perspective. Even if a trait is perfectly stable, stability coefficients will be limited by the reliability of the measures of that trait: "In the interpretation of correlations from age to age and from one ability to another, allowance must be made for errors of measurement: errors in the total score, in part scores, in subtest scores, and in the differences between these" (Cronbach, 1984, p. 235). In most of the research that will be presented, reliability estimates are not available for the samples studied. Where possible, reliability estimates for the same or similar measures from other samples will be presented to aid in the interpretation of the obtained stability coefficients. Since reliability estimates for a given test often differ between samples, these reliabilities will serve only as rough estimates of the reliabilities for the samples in question.

Although longitudinal studies are appropriate for estimating stabilities, some biases inherent in longitudinal data should be kept in mind. Longitudinal samples are often not very representative of the general population. One reason for this is differential attrition over the course of the study. There are often systematic differences between those who leave and those who stay. Many longitudinal samples represent a restricted range of talent, in part due to the differential attrition, but also because high ability samples have more often been studied. Finally, repeated measurement of the same subjects over time might have an effect on their test scores.

Cognitive Abilities

A substantial amount of research is available on the stability of cognitive abilities over time and on the stability of general intelligence as well. Tests of general intelligence are very similar to cognitive ability tests, the difference being that intelligence tests draw their items from an even wider range of human experience than cognitive ability tests (Angoff, 1988). Jensen (1981) goes one step further and says that cognitive ability tests measure general intelligence to about the same degree as intelligence tests, and that the two types of tests are more or less functionally equivalent. Tests of general intelligence are very similar to cognitive ability tests at the least, so we can expect the stability of these two types of tests to be very similar. Research on the stability of general intelligence will be included in the present review. Cronbach (1984) estimates that the immediate test-retest reliabilities of individually administered intelligence test scores are in the low .90s, and that the reliability of group administered intelligence tests is about the same. Reliabilities will obviously vary a great deal between different tests, but this provides a context in which to interpret some of the stabilities presented in this section, for which the appropriate reliabilities are not available.

One extensive study of the stability of general intelligence in childhood and adolescence involved retesting a sample of 252 children on individually administered intelligence tests many times over the course of 18 years (Honzik, MacFarlane, & Allen, 1948). The first administration was shortly after birth and the last was at age 18 years. Results show that intelligence test scores become more stable with age and that, after about age 8, intelligence test scores are very stable. Other studies have reported similar findings. By about age 14, intelligence test scores and scores on most tests of more specific cognitive abilities are very stable, and, by age 18, these scores are extremely stable. In addition, Honzik et al. found that the correlation between tests administered closer together in time is consistently higher than for those administered further apart in time. For tests administered after the age of 8, the average correlation between tests administered one year apart is .90, while the average correlation between tests administered two years apart is .88. Intercorrelations of cognitive abilities over time often display this pattern (e.g., Henry & Hulin, 1987). This pattern of correlations has been called a simplex or quasi-simplex structure.

In a high school sample, stability coefficients for the cognitive abilities measured by the General Aptitude Test Battery (GATB) (U.S. Department of Labor, 1970) appear to fit the simplex as well (Droege, 1966b). The GATB measures four cognitive abilities: intelligence, verbal aptitude, numerical aptitude, and spatial aptitude. It was administered to large samples of high school freshman, sophomores, and juniors (about 7000 each). All three groups were retested as seniors. For the juniors (retested after one year), stability coefficients range from .74 for spatial aptitude to .83 for general intelligence. For the sophomores (retested after two years), stabilities range from .73 to .82, and for the freshmen (retested after three years), they range from .70 to .79. There is a decrease in stability coefficients for longer time periods, but, since age at initial testing is confounded with the length of retest interval, it is impossible to tell whether the differences in stabilities are due to differences in the time interval, the instability of scores at younger ages, or a combination of the two. In the GATB manual, test-retest reliabilities for high school seniors over a three-month interval are reported to range from the low .70s to the mid .80s.

A study of the stability of the GATB in an adult sample did not find a simplex pattern. Droege (1966a) administered the GATB to a sample of about 900 adults (average age, about 30 years), then

divided them into three subsamples. Each subsample was retested with an alternate form of the GATB after one, two, or three years. Scores for all four cognitive abilities are very stable, with stability coefficients after one year ranging from .83 to .90; after two years, from .79 to .87; and after three years, from .84 to .90. These stabilities are slightly higher than those found in the high school sample. In addition, there is no apparent decrease in stability over the time periods studied. The median stability after one year is .84; after two years, .86; and after three years, .87. It is not clear why a simplex pattern was not found in this study, since so many other studies have found simplex patterns. The fact that the different time periods also represent different samples could account for this difference, but the samples were not reported to differ in any systematic way. Test-retest data for adult samples presented in the GATB manual provide a context for interpreting these stability coefficients. These reliabilities are generally in the high .80s to low .90s. When the obtained stability coefficients are contrasted with this typical level of reliability, it appears that these abilities are extremely stable. Since the ages of most military applicants are likely to lie between those of the high school and adult groups that have been studied using the GATB, the stability of these abilities would probably lie somewhere between those reported here for high school students and adults.

Long term longitudinal studies have also reported high stabilities for most cognitive abilities, even over periods as long as 30 years. Since the samples studied often represent a restricted range of talent, these stabilities may actually be underestimates of the true stability in the entire population. Eichorn, Hunt, and Honzik (1981) combined data from three studies and found an overall correlation between intelligence test scores in late adolescence and scores in middle age (average test-retest interval about 24 years) to be .83 for a total of 117 males and .77 for a total of 133 females.

In another study, a sample of 127 college freshmen was tested on the Army Alpha, a group administered intelligence test. When they were retested 31 years later, the stability coefficient for total score was .77 (Owens, 1966). When this same sample was retested a second time after 11 more years, the 42-year stability was .78. The stability over the 11 years between these two retests is .92. Since the author reports that this sample represents a restricted range of talent, these could be underestimates of the long term stability of Army Alpha scores.

For a sample of 110 young adults who were retested on an individually administered intelligence test after 14 years (average age about 14 at first testing), the correlation between scores from the two administrations is .85 (Kangas & Bradway, 1971). In yet another study, a sample of 164 men were retested on the Army General Classification Test (AGCT) after a 13-year interval (average age, about 30 at first testing); the correlation between scores from the two administrations is .79 (Tuddenham, Blumenkrantz, & Wilkin, 1968). Finally, over an interval of about 10 years, the stability of scores on the Concept Mastery test is .90 (Bayley & Oden, 1955).

This brief review shows that general intelligence and most cognitive abilities are extremely stable, particularly over time periods as short as one to three years. Humphreys (1986) estimates that, for general intelligence, the correlation between true scores over a single year (after correcting for the unreliability of the tests) is in the high .90s. The present review supports this conclusion.

Some cognitive abilities, however, are more stable than others. For example, in tests of general intelligence, the verbal subscore has consistently been found to be more stable than the

performance subscore (e.g., Kangas & Bradway, 1971). Scores on tests that are thought to measure "crystallized" intelligence, such as vocabulary and mathematics, are more stable than scores on tests that are thought to measure "fluid" intelligence, such as reasoning or memory span (Vernon, 1979).

The stabilities of scores on the ASVAB tests and composites are likely to be in the same range as the stabilities of the cognitive abilities reviewed here, because some of the abilities measured by the tests in the present review are similar to those measured by tests in the ASVAB (e.g., verbal ability). In addition, reliability estimates for the ASVAB composites are comparable to those for tests reviewed above. For example, internal consistency estimates for the ASVAB composites range from .88 to .92 (*Counselor's Manual*, 1984). Test-retest values over a two-week interval using alternate forms of the ASVAB range from .84 to .95 (*Counselor's Manual*, 1984). Finally, stability coefficients for the ASVAB tests have been reported by Christal (1989) for groups of Air Force airmen who were retested on the ASVAB after time periods of 0-6 months (N = 1174), 6-12 months (N = 1785), and more than 12 months (N = 518). For speeded ASVAB tests (Coding Speed and Numerical Operations), the reported stability coefficients are about .69 for the 0-6 month group, .66 for the 6-12 month group, and .61 for the more-than-12-month group. For the remaining ASVAB tests (not including Paragraph Comprehension), which are primarily power tests, the 0-6 month stabilities range from .69 to .84, the 6-12 month stabilities range from .69 to .82, and the more-than-12-month stabilities range from .66 to .82. Paragraph Comprehension has much more lower stabilities, with stability coefficients of .49, .41, and .35 respectively for these same three groups. The stability coefficients for the various ASVAB tests differ quite a bit, and it is likely that the stabilities of the various ASVAB composites will differ as well. In addition, the ASVAB composites are likely to be even more stable than the individual ASVAB tests, because each composite is composed of several ASVAB tests. Based on the present review, true score correlations over one year for these composites are likely to be in the middle to high .90s.

Perceptual and Psychomotor Abilities

There is not a great deal of research on the stability of perceptual and psychomotor abilities but available research indicates that most of these abilities are very stable over time periods of up to three years. In general, perceptual and psychomotor abilities do not appear to be quite as stable as cognitive abilities, and there are larger differences in stability among the various perceptual and psychomotor abilities.

The General Aptitude Test Battery (GATB) contains tests of two perceptual abilities and three psychomotor abilities. The two studies of the GATB described in the cognitive ability section provide information on the stability of perceptual and psychomotor abilities as well. For the adult sample, stability coefficients for the five perceptual and psychomotor abilities after one year range from .74 to .85; after two years, from .69 to .85; and after three years, from .73 to .84 (Droege, 1966a). The stability coefficient for motor coordination is consistently a little higher than those for the other abilities. Note that, in this adult sample, stabilities again do not appear to decrease as a function of time. The median stability after one year is .76; after two years, .74; and after three years, .76.

For the high school group, stabilities for these five abilities range from .64 to .79 after one year, from .62 to .74 after two years, and from .57 to .70 after three years (Droege, 1966b). As with the cognitive abilities discussed earlier, it is difficult to say how much of this decrease in reliability for

longer time periods is due to the fact that age at initial testing is confounded with length of retest period and how much is really due to instability over the longer period of time. In the GATB manual (U.S. Department of Labor, 1970), the two-week test-retest reliabilities for these abilities in adult samples range from the low .80s to the low .90s. The manual, also reports the three-month test-retest reliabilities for high school students, and these range from the high .60s to the low .80s with most in the middle .70s. Motor coordination is consistently one of the most reliable of the perceptual and psychomotor tests in all samples.

A few perceptual and psychomotor abilities appear to be very unstable, even over short periods of time. Peterson (1987) reports two-week test-retest reliabilities for a group of experimental perceptual and psychomotor tests for a sample of about 120 soldiers. Even though the split half reliabilities for all of these scores are high (ranging from .86 to .97 for the final group of scores), some of these scores have very low test-retest reliabilities, even over this short time period. For example, the test-retest reliability for simple reaction time is only .37; for choice reaction time, .56; and for the time to fire score on a target shoot test, only .48. The test-retest reliabilities for other perceptual and psychomotor tests in this battery, however, are generally higher (in the .60s and .70s).

Personality Traits

It is difficult to summarize the literature on the stability of personality traits in a meaningful way. A wide variety of personality traits have been studied, and multiple instruments have been developed to measure many of them. In addition, some measures that have different labels actually measure similar traits, and some measures with the same label are actually not very similar.

Another problem in interpreting stability coefficients from the available research is that some scores on personality instruments are not very reliable. Some personality variables are actually expected to change from one administration to another (e.g., mood), while others such as "temperaments" are expected to remain stable over the entire life span. Short term test-retest reliabilities for personality scales tend to be lower than those for ability tests, and they vary a great deal from one inventory to another. For the California Psychological Inventory (Gough, 1975), reliabilities over about a month for an adult sample are in the .70s and .80s. For some scales and inventories, however, reliabilities can be much lower. For example, the mean test-retest reliability for the 16PF was reported to be only around .50 (Siegler, George, & Okun, 1979). Any stability coefficients for personality scales will obviously be limited by the test-retest reliabilities. Since many personality scales have only moderate reliabilities and since estimates of the reliabilities are usually not available for the samples on which the stability coefficients are computed, interpretation of many of the results is difficult.

The present review will focus on paper-and-pencil personality inventories, since this is the type of personality measure that is likely to be used in personnel selection. The stability of paper-and-pencil types of personality measures has not been studied for children and adolescents, but research using other personality measures (e.g., ratings) typically shows that personality traits are less stable for children and adolescents than for adults. Moss and Susman (1980) conclude that, when studies include both children and young adults, the stability coefficients are always higher for the older group.

The available evidence on the stability of personality in adults indicates that most personality traits, as measured by paper-and-pencil personality inventories, are moderately stable over time.

There is some evidence, however, that the stability of some personality traits is extremely high once correction has been made for the unreliability of the measures. The average stability of scores on the 16PF scales over a two-year interval for a sample of adults is .50 (Siegler et al., 1979). Two-week reliabilities for this same sample are at exactly the same level as these stabilities. Note that these scales do not have particularly high reliabilities to begin with, but the reliable variance appears to be very stable. Similar stabilities were found for a sample of adult males (Costa & McCrae, 1978). In this second study, the measured traits are categorized according to three higher order dimensions of personality: neuroticism, extroversion, and openness. Stabilities for scales measuring neuroticism range from .58 to .69; for those measuring extroversion, from .70 to .84; and for those measuring openness, from .44 to .63. Extroversion seems to be the most stable. However, scales measuring neuroticism are less reliable, and these scales are also quite stable when corrected for unreliability. Other researchers have found similar results for these higher order dimensions.

In a sample of adult males, the stabilities over six years for scale scores on the Guilford-Zimmerman Temperament Survey range from .68 to .86 with a mean of .77 (Costa, McCrae, & Arenberg, 1980). Over 12 years, the stabilities for this same sample range from .61 to .87 with a mean of .73. When the sample is divided into three age groups (17-44, 45-59, 60-85), there are no differences in the stability coefficients among these three age groups. Scales related to extroversion and neuroticism are again more stable than other scales (when corrected for unreliability). The one-year stability coefficients cited in the manual (Guilford, Zimmerman, & Guilford, 1976) for these inventories are actually slightly lower than the stability coefficients obtained in this study.

For a sample of junior college students, the median stability for scale scores on the California Psychological Inventory (CPI) over an interval of about two years is about .60 (Plant & Telford, 1966). Short term reliability estimates for these scales for adults are in the .70s and .80s, but reliabilities for a junior college sample may differ from those for adults. The CPI manual (Gough, 1975) also reports one-year stabilities for high school students ranging from the low .40s to the .70s with a median of about .60.

No evidence was found that dramatic changes in life style result in personality change. Subjects in the studies reported here are likely to have had a wide variety of life experiences between the two administrations of the personality inventory. It is likely that the experiences between the two administrations do not differ in any systematic way from the rest of their life experiences. Military training, on the other hand, is likely to differ dramatically from previous life experiences. It involves a totally different environment and a conscious attempt by the military organization to change certain behaviors. Literature could not be found on the impact of this or other such total reorganizations of individuals' day-to-day environments on the stability of personality variables. If the military environment does have an impact on the traits measured by personality inventories, this could lower the stability of these variables.

Interests

Scores on interest inventories are very stable for adults. Patterns of interest scale scores are even more stable than the scale scores themselves. Interests are more stable in adults than they are in children and adolescents. They become quite stable by about age 18, and they are very stable after about age 22.

There are several studies available on the stability of scores on the Strong-Campbell Interest Inventory. The manual for the Strong-Campbell reports median test-retest reliabilities in adult samples of .91 for a two-week interval, and .89 for a 30-day interval (Campbell & Hansen, 1981). The manual reports a median stability for these scales over three years of .83. Hansen and Stocco (1980) report a median scale stability over four years for a sample of young adults who were originally tested as college freshmen of .67. They report stabilities about ten points lower for a sample of adolescents (first tested as high school freshmen) over a similar time period. For the college sample, Hansen and Swanson (1983) report correlations between profiles of scores at the two administrations ranging from -.23 to .97 with a median of .78. For a group of men first tested at age 15, Strong (1943) reports a median stability over 10 years of .82; after 40 years, the median stability for this same group is .73. The correlation between the latter two administrations (a 30-year interval) is .88.

For the Kuder Occupational Interest Inventory, similar results have been obtained. For a sample of university students, the median stability of patterns of interests over about three years is .89, ranging from .35 to .98 (Kuder & Diamond, 1971). Two-week test-retest reliabilities for this inventory based on other samples range from .90 to .96.

The Milwaukee Academic Interest Inventory has 150 items that are combined into nine interest scales. Two-year stabilities for the scale scores in a sample of adolescents range from .72 to .87 with a median stability of .82 (Baggaley, 1974).

For a sample of about 300 young men in the Navy, the median profile stability over about four years for scores on the Navy Vocational Interest Inventory is .86 (Dann & Abrahams, 1973). Stabilities over six years for the individual scales for a sample of non-reenlistees range from .45 to .79 with a median of .62. For a sample of reenlistees, the stabilities are slightly higher with a median stability over five years of .70. The median internal consistency reliability for these scales is .96.

Biographical Questionnaires

Little data are available on the stability of scores from biographical questionnaires. The few studies that have been done have found that these scores are very stable over time. Since biographical questionnaires or "biodata" inventories are typically used in a selection context, research is not available on the stability of scores for children and adolescents. Test-retest reliabilities for the instruments used in the studies described below were not reported, but for other biodata instruments test-retest reliabilities obtained over a few days or weeks have been in the low .90s.

Mumford and Owens (1982) report the six-month stabilities of biodata scores for a sample of about 400 young adults who were first tested as college freshmen. A 389-item biodata inventory was administered to this sample, and factor scores for males and females were derived separately. For males, there are 13 factor scores, and stabilities range from .91 to .97. For females, there are 15 factor scores, and stabilities range from .77 to .97.

Shaffer, Saunders, and Owens (1986) conducted a similar study with a shorter scale and a longer test-retest period. Their sample of 237 young adults was first tested when they were college freshman and retested five years later. The inventory was an 118-item biographical data

questionnaire containing many of the same items as the previous study. For males, stabilities for 13-factor scores range from .49 to .91 with a median of .74. For females, stabilities for 14-factor scores range from .50 to .88 with a median of .74. Based on these two studies, it appears that stabilities for many of the scales are lower for longer retest intervals. However, since the study with the longer retest interval also uses shorter scales, the lower stabilities for the longer time interval could be partly due to the shorter scales.

Because biographical questionnaires ask for information concerning events from the respondent's past, it is not clear whether information collected closer in time to the criterion of interest would be expected to be a more or less valid predictor of that criterion. An argument could be made that biodata questions asked earlier in time are more likely to be based on accurate memories of the events, since memory deteriorates over time. Based on this line of reasoning, we expect the earlier administration to be more valid than the administration closer to the criterion of interest. An argument could also be made, however, that the variable of interest is actually the respondent's perception of these past events. Based on this line of reasoning, the closer the administration of the biodata instrument is to the criterion, the more likely it is that these perceptions will be similar at the two times. The later administration would then be expected to be more valid. This question cannot be answered based on the available data, but biodata are the only individual differences measures reviewed here for which there are some reasons to believe that scores collected further in time from the criterion might actually be more valid.

Conclusions About the Stability of Psychological Traits

Cognitive abilities are extremely stable in 18- to 25-year-old men and women. There are some differences between the stabilities of various cognitive abilities, but for most abilities, the true score correlations over a one-year interval are probably in the high .90s. Most research has found that stability coefficients for longer time periods are lower, and true score correlations over a two- to three-year interval are probably in the middle to low .90s. Perceptual and psychomotor abilities are slightly less reliable than cognitive abilities, and stabilities over a one-year interval for most of these abilities are in the .70s and .80s. However, some perceptual and psychomotor abilities, for example, simple reaction time, are very unstable even over short periods of time. All abilities are less stable for high school students than for adults, but by age 18 stabilities are very high.

Measures of personality traits are generally less reliable than ability tests, and stability coefficients are correspondingly lower. However, many personality traits appear to be quite stable when corrected for unreliability, and some appear to be extremely stable. Because there are problems in correcting for unreliability and because results for different traits are often very different, it is difficult to make any generalizations about the level of stability that might be expected for scores from personality inventories.

Interests are very stable after about the age of 22, but stabilities for younger samples (ages 18 to 22) are slightly lower. Stabilities over three-year intervals in adult samples are in the low .80s, but for the same time interval, stabilities in a sample of college students are in the .60s. Stabilities are even lower for high school samples. Patterns of interest scale scores are more stable than the individual scales, but individual correlations for these patterns range widely.

There is not much evidence concerning the stability of biodata, but the available evidence suggests that scores from biographical questionnaires are very stable, with stabilities over six months reported in the .90s and stabilities over five years reported in the .70s.

Implications for the Research Question

The TC hypothesis cannot account for changes in the validity of cognitive abilities measured by the ASVAB over time unless those abilities actually change over time. In other words, TC will not even be plausible unless the rank order of true scores changes over time. Based on the present review, changes in the rank order of scores over time for cognitive abilities are negligible. True score correlations between two administrations (the stability coefficients corrected for unreliability) are typically very high. For a one- to two-year interval they are likely to be in the mid to high .90s for most cognitive abilities. Unless the observed changes in validity over time are quite small, TC alone cannot account for these changes. It is possible, however, that intervening events could lead to larger changes in individual differences and consequently lower stabilities.

The ASVAB is not purely a measure of cognitive abilities. Some of the ASVAB subtests can be seen as having a large "achievement" or "knowledge" component (e.g., Electronic Information and General Science). Scores on these tests could be expected to change more over time than scores on the other ASVAB subtests, particularly for applicants who are still in school. Christal (1989) presents some evidence that tests measuring "technical knowledge" (Electronic Information, Mechanical Comprehension, General Science, and Auto-Shop Information) are in fact slightly more valid when they are administered closer in time to the criterion measure. Specifically, when the ASVAB was readministered to random samples of Airmen on the sixth day of Basic Military Training, the second administration of these four "technical knowledge" subtests is generally a more valid predictor of Technical School grades than the original administration. For the remaining tests, classified as measuring "general ability" and "perceptual speed," there are no consistent differences in the validity of the two administrations. These results suggest that TC may be a better explanation of changes in validity over time for some ASVAB subtests than for others, but since TC and BAE are confounded in the Christal study this conclusion is only tentative.

The Effect of Intervening Events on Test Scores

The purpose of this section is to discuss potential sources of change in test scores and to estimate the potential magnitude of changes due to each source. Change can be unintentional, as in maturation, or it can be intentional, as in schooling, training, and other learning experiences. There can also be changes in test scores without corresponding changes in the underlying abilities; for example, changes due to practice, coaching, or improved test taking skills. In addition, the motivation to do well, the desire to portray oneself in a certain manner, and levels of anxiety about the testing situation may differ between two administrations and lead to increases or decreases in test scores. Changes in test scores due to these different sources will have different implications for changes in the validity of these scores.

Maturation

Test scores for most cognitive abilities increase rapidly until about age 14, when the rate of increase slows a great deal. For many abilities, scores continue to increase slowly through the late teens and early twenties. For some abilities, scores continue to increase throughout much of the life span (e.g., Vernon, 1979). For example, some verbal abilities (e.g., vocabulary) continue to increase throughout adulthood, and other abilities including reasoning and numeric ability have stable means starting in the late teens or early twenties (e.g., Owens, 1966). Gains continuing into late adulthood are typically found for tests that measure "crystallized" intelligence

(e.g., information and vocabulary tests). Declines by about middle age are typically found on tests where speed of motor response is important. The results for "fluid" intelligence have been mixed (Eichorn et al., 1981). General intelligence appears to peak in the late twenties in cross-sectional studies, but when studied longitudinally scores continue to increase beyond the age of 40. Even though there is some growth later in life for many abilities, the rank order of individuals' scores is generally very stable by the early twenties.

Droege (1966b) studied the effects of maturation on the perceptual and psychomotor abilities measured by the GATB in a high school sample. Scores were found to increase as a function of age throughout high school, but this increase has slowed by the twelfth grade.

For interests, Strong has estimated informally that the change between the ages of 15 and 25 can be divided into thirds. The first third occurs between ages 15 and 16; the second, between ages 16 and 18; and the last, between ages 18 and 25 (Campbell, 1971).

We conclude from all this evidence that maturation should have little effect on the stability of test scores over time intervals from one to three years for 18- to 25-year-old men and women. Although many abilities continue to increase throughout much of the life span, the changes after about age 14 are very small and the rank order of individuals remains very stable. Rapid changes in abilities due to maturation are not likely to be occurring in the military applicant population.

Schooling, Training, and Other Life Experiences

Several studies have concluded that schooling has a substantial effect on intelligence test scores. There is little evidence, however, that the type of school or the program of study has much impact on cognitive ability or intelligence test scores (Vernon, 1979). Most of the research in this area has focused on children. For example, Wheeler (1942) attributes differences as large as one standard deviation between children tested in 1930 and those tested in 1940 to the substantial improvement in living conditions and education that had occurred between the two testings. Husen (1951) and Lorge (1945) both found that young adults who had completed secondary school obtained intelligence test scores at age 20 that are about a standard deviation higher than those who did not complete secondary school, even after controlling for initial test scores.

Vernon (1957) found that the type of schooling (grammar school vs. modern school) has a moderate (seven point) effect on intelligence measured three years later, after controlling for initial scores. Many other studies, however, have failed to find a substantial effect for the type or quality of school on abilities or general intelligence. For example, Angoff and Johnson (1988) grouped a large sample of college students into four groups based on their field of study, and they compared their scores on the SAT administered prior to college with scores on a similar test, the GRE, administered after college. They found that scores on the two tests (administered about four years apart) are highly correlated, with correlations in the mid .80s. In addition, they found that the impact of field of study on the verbal score is low, and on the quantitative score, the effect is moderate. The institution attended has no impact on the verbal score and only a slight effect on the quantitative score. They report that this is typical of studies in late adolescence; that scores on ability tests at this age change little in response to intervention.

Several studies have looked, post hoc, at the relationship of gains or losses in intelligence test scores over time with various life experiences and personality variables. Correlations of these personality variables and life experiences with changes in test scores are generally low and often

not significantly different from zero. Results from analyses on extreme groups of increases and decreases, however, often show significant effects. For example, one study found that education level attained, parents' education, father's occupation, spouse's intelligence, and scores on several CPI scales (intellectual efficiency, achievement via independence, tolerance for ambiguity, responsibility, capacity for status) are related to the amount of increase or decrease in intelligence test scores over a period of about 24 years (Eichorn et al., 1981). In addition, poor physical health, problem drinking, depression, and lack of stimulation are related to extreme decreases, and traveling outside the country and having a more intelligent spouse are related to extreme increases. Other studies have also found significant relationships, but not exactly the same pattern as reported here (e.g., Owens, 1966).

Although schooling can have a fairly substantial impact on ability test scores, it is not likely to have a large effect on the stability or validity of test scores for military applicants for several reasons. First, schooling has only been shown to have large effects on children's test scores. Abilities are extremely difficult to change after childhood. Angoff (1988, p. 719) states that "although it is quite well understood, it will bear repeating that changes in the individual are much more easily effected if efforts to make them are instituted in early childhood. They are not, however, easily effected if the efforts to make them are delayed until adolescence or later." Second, changes in abilities have generally not been shown to be related to type of schooling. Since all military applicants are immediately given some form of schooling, any increases in ability that do occur will affect the entire sample. Unless these changes are large, they will probably only have a minor effect on the rank order of scores or the stability coefficient.

Changes in Test Sophistication or "Test-Wiseness" and Test Anxiety

Test sophistication or "test-wiseness" has been defined by Millman, Bishop, and Ebel (1965) as "a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score." Their definition includes a variety of test taking strategies that apply to nearly all tests, but excludes other factors such as test anxiety, motivation, and preparation for the test. They provide an analysis of the components of test-wiseness, which can be broken down into two broad categories; those components that are independent of the test constructor or test purpose, and those that are dependent on the constructor or purpose. The former include time-using strategies, error-avoidance strategies, guessing strategies, and deductive reasoning strategies. The latter include intent consideration strategies and cue using strategies. This definition seems widely accepted, and research has shown that test-wiseness, as defined here, is related to higher scores on ability and achievement tests (Vernon, 1979). In addition, researchers have successfully taught test-taking strategies to children, and this training has been found to result in higher scores on ability and achievement tests (e.g., Wahlstrom & Boersman, 1968).

Measures of test-wiseness have been found to correlate with age for elementary school children, but beyond elementary school these scores are not related to age (Vernon, 1979). It appears that after elementary school most children, at least in the United States, have already mastered test taking strategies. Children in the United States are almost always introduced to objective tests early in their school careers, and they are likely to be thoroughly exposed to tests early in grade school. Therefore, large increases in test-wiseness are not likely to occur later in life.

Test anxiety has also been found to affect test scores; examinees with high levels of test anxiety obtain lower scores. For those who experience test anxiety, greater familiarity with the test taking situation can lower this anxiety and improve test scores (Anastasi, 1976).

Changes in general test-wiseness or test sophistication are only likely to have an impact on the present research question if a substantial proportion of the examinees are initially very inexperienced in taking multiple choice tests. Similarly, a second administration of the ASVAB would only be expected to result in lower test anxiety if individuals are very unfamiliar with the test taking situations. For the large majority of applicants, this is highly unlikely for most of the tests in the ASVAB, since people in the United States are typically exposed to taking this sort of test at an early age. Therefore, changes in test-wiseness or test anxiety are unlikely to have an impact on most ASVAB test scores. Some groups of applicants, however, may be less exposed to testing situations or be much more "test anxious." Generally, these would be groups less in the mainstream of the American educational systems. For such groups, efforts at increasing test sophistication and decreasing test anxiety may pay dividends in terms of increased scores. As noted, the impact on validity of the tests would depend on several factors, including the size of these groups. It is possible that some of the tests in the ASVAB will be more novel to the general population than others. Several ASVAB tests (Coding Speed, Electronic Information, and Mechanical Comprehension) appear to be relatively dissimilar to the ability tests that are commonly used in the American educational system. However, the test-retest reliabilities for these tests are comparable to those for other ASVAB tests of the same type (i.e., speeded or power), and they show very little increase in mean scores on the second testing (Christal,1989). Thus, there do not appear to be substantial shifts in test-wiseness or test anxiety for a large proportion of examinees, even on these tests.

Practice Effects

Retests have been found to lead to consistent gains in cognitive, perceptual, and psychomotor abilities, and sometimes these gains are fairly large. Increases with practice have been found in a wide variety of groups, but different abilities appear to be more or less susceptible to practice effects. For example, repetition of individually administered intelligence tests (e.g., the Terman-Merrill or the WISC) within less than one year leads to a distinct practice rise, especially on the performance subtests (Vernon, 1979). Cronbach (1984) reports average retest gains on another individually administered intelligence test, the WISC-R, of one fifth of a standard deviation for verbal scores and three fifths of a standard deviation for performance scores. For the abilities measured by the GATB, practice effects have been consistently found in several studies; for some tests, these are fairly large, almost half a standard deviation (Droege, 1966b). GATB tests with the largest practice effects include manual dexterity and those with the smallest include spatial perception and numerical aptitude. A group of experimental spatial tests developed for the Army (Peterson, 1987) had practice effects ranging from near zero to three fifths of a standard deviation for a test-retest interval of 2 weeks. In this same study, smaller practice effects are reported for several perceptual and psychomotor tests (the median effect is about one tenth of a standard deviation). Practice is often a component of coaching programs, and research that involves both practice and other types of coaching will be discussed in the following section on coaching as will the implications of practice effects for the research question.

Coaching

Coaching involves a deliberate, relatively short term effort to raise test scores. Vernon (1979) says that coaching, even on intelligence tests, is widespread. Coaching is "instructions given in preparation for taking a test that are designed to elicit maximum performance on the part of the individual coached" (Cole, 1982). Coaching can take many different forms. Since it often involves

practice on similar items or instruction in test taking strategies, the literature on practice effects and test-wiseness is relevant. Simple practice, however, is often used as the control against which the effects of coaching are compared.

Cole (1982) proposed the following taxonomy of the components of coaching:

a. Supplying correct answers (cheating)
b. Taking the test for practice
c. Maximizing motivation
d. Optimizing test anxiety
e. Instructing test-wiseness
f. Instructing test content.

This taxonomy outlines all possible approaches to increasing test scores in a relatively short time period, and a given coaching program will probably include only a subset of these components.

It should be noted that the category labeled "instructing test content" is not meant to include formal schooling or training. Although schooling could also be thought of as a conscious effort to raise test scores, coaching is by definition of shorter duration. Coaching is not typically expected to actually change the underlying knowledge, skill, or ability that is being measured. In addition, since coaching involves a concentrated effort to increase test scores in a short period of time, coaching will probably result in the maximum change in test scores possible without corresponding changes in true scores.

The effects of coaching on ability tests vary greatly from study to study. The several literature reviews that have been done draw somewhat different conclusions. When these reviews have attempted to estimate the magnitude of the effect of coaching on ability test scores, the resulting estimates have differed. For example, one review of the effect of coaching on SAT scores estimated that the average effect size for coaching is only about one tenth of a standard deviation (DerSimonian & Laird, 1983). A similar review (Kulik, Bangert-Drowns, & Kulik, 1984) resulted in slightly higher estimates. A review of the effect of coaching on other aptitude and intelligence tests (Messick & Jungeblut, 1981) found that coaching raises scores an average of .43 standard deviations. One possible reason for the inconsistent results found for coaching is that different coaching programs employ different subsets of the components of coaching presented in the taxonomy above. In addition, some of the coaching programs reviewed have been much longer in duration than others. There is mixed evidence concerning whether there is a larger effect for longer coaching programs, but the most careful analyses have not found this effect. Some researchers hypothesize that the size of the effect depends on the type of coaching; the more similar the coaching materials are to the test items, the greater is the effect (Yates, James, Dempster, Wiseman, & Vernon, 1953).

Another reason for the inconsistent results found for coaching programs is that some tests and some item types seem to be more susceptible to coaching and practice than others. A meta-analysis by Powers (1986) found that some item formats are more susceptible to practice and coaching effects than others. In addition, tests with longer, more complex instructions were also found to be more susceptible to coaching. Additional evidence that different tests have different susceptibilities to coaching is found in another meta-analysis (Kulik et al., 1984), where they found smaller coaching effects for the SAT than for the other ability tests studied. Since the SAT is designed to minimize the effect of coaching on test scores, this is not too surprising.

Most researchers seem to consider coaching a threat to the validity of test scores. Cole (1982) discusses several ways in which coaching can affect validity, all of which are expected to result in lowered validity. For example, if coaching raises test scores above the level of the individuals' true abilities, these scores would no longer be valid measures of their abilities. Even if coaching is assumed to help examinees achieve the maximum test score possible given their ability level, coaching that is available to some examinees and not others would be a source of bias in test scores and therefore likely to have adverse effects on validity. One provocative line of research (Embretson, 1987) has found that changes in test scores after a very specific coaching program actually led to scores that are more internally consistent and more valid in predicting an external criterion. If replicated by other studies, this finding suggests that coaching, if available to all examinees, would actually have a positive effect on validity. In this study, coaching could be seen as actually altering the nature of the test and the abilities tested. Another study also found that the factor composition of a test frequently depends on the problem solving styles used in answering the test items (French, 1965).

For most standardized ability tests, coaching has not typically been found to result in large changes, and practice effects are generally even smaller. However, these effects have been consistently in a positive direction, and Embretson (1987) raised test scores quite a bit (about two thirds of a standard deviation) by using very specific coaching. Other studies often use less specific coaching and have generally found smaller effect sizes. The SAT shows the smallest effect sizes, and several plausible reasons for this will be briefly discussed. First, the SAT items have been designed to minimize coaching effects. Secondly, these items have not been available, until recently, to those who design the coaching programs. Finally, it is conceivable that tests that contain item content sampled from a large and diverse domain will be much more difficult to coach than will tests of more narrow abilities that contain relatively homogeneous item types (e.g., special aptitude tests such as tests of spatial ability). The SAT samples from virtually all high school subjects, and thus would be expected to be more difficult to coach than many other tests.

Changes in Levels of Motivation and Other Temporary States

Although examinees' motivation to do well on a test or to portray themselves in a certain manner can be very different between two test administrations, these differences would have to affect a fairly substantial portion of the examinees before an effect on the validity of test scores will be noticeable. In addition, if the levels of motivation differed randomly among the examinees between two occasions, the validity of scores from the two occasions would be expected to be the same. We would only expect changes in validities if there is reason to believe that motivation levels for many examinees differ systematically between the two testings.

Systematic differences could occur if one of the administrations is to be used in selection and the other is for research only. For example, we might expect respondents in a selection situation to try to portray themselves in a positive way on personality and interest inventories. Research has found that, when asked to, people can fake personality and interest inventories, but it appears that they do not do so in a selection setting (Peterson, 1987). We might also expect respondents to try harder on an ability test if it is administered in a selection context. However, for the present research question, this would be expected to lead to higher rather than lower validities for the earlier administration of the tests.

Implications for the Research Question

We conclude from all this evidence that maturation should have little effect on the stability of scores on individual differences measures over time intervals from one to three years for 18- to 25-year-old men and women except, perhaps, for vocational interests. Rapid changes in cognitive abilities due to maturation, especially, are not likely to be occurring in the military applicant population. Although schooling can have a fairly substantial impact on ability test scores, it is not likely to have a large effect on the stability or validity of test scores for military applicants for several reasons. Schooling has only been shown to have large effects on children's test scores. Abilities are extremely difficult to change after childhood. Also, changes in abilities have generally not been shown to be related to type of schooling. Thus, although some military applicants may have an additional year of high school before induction into the Armed Services, the abilities measured by the ASVAB may not change much. Changes in test-wiseness, test sophistication, or test anxiety may have an impact on the present research question, since some ASVAB tests have relatively novel content, but there is presently little evidence to support this point of view.

For most standardized ability tests, coaching has not typically been found to result in large changes, and practice effects are generally even smaller. These effects have, however, been consistently in a positive direction. One interesting hypothesis is that coaching and practice will actually lead to test scores that are more valid predictors of external criteria. Very specific coaching, more general coaching, and simple practice could be seen as lying on a continuum. If the Embretson (1987) finding of gains in validity with specific coaching holds up in future studies, we might expect gains in validity for less specific coaching and practice as well (assuming the treatment is the same for all examinees), but probably smaller gains than those found by Embretson.

Decreases in Validity Over Time

Findings From College Student Samples

Humphreys and other researchers (Angoff & Johnson, 1988; Humphreys, 1960, 1968; Humphreys, Park, & Parsons, 1979; Humphreys & Taber, 1973; Lin & Humphreys, 1977; Wilson, 1983) have presented convincing and consistent evidence that aptitude tests (like the Scholastic Aptitude Test) show decreasing validity over time for predicting academic performance in college. The basic methodology followed in these studies is to correlate scores on aptitude tests with annual, independently computed grade point averages (GPA). The aptitude test scores were obtained from test administrations prior to admission to undergraduate school, and in some of the studies, from test administrations during the senior year (usually for purposes of admission to graduate school). Correlations of aptitude test scores with GPA show a decline over the four undergraduate years.

The independently computed, annual GPAs conform to the simplex or quasi-simplex form that we described earlier (that is, correlations between scores for adjacent time periods are high and constant [or increasing] over time, whereas correlations between scores for non-adjacent time periods are lower and decrease with distance in time). Furthermore, the decline in validity coefficients from freshman to senior years is present for scores from aptitude tests administered during the senior year as well as for tests administered prior to the freshman year. Finally, the correlations between aptitude test scores administered prior to the freshman year and during the

senior year appear to be very high (Angoff & Johnson, 1988, report correlations of .86 for both SAT Verbal-Graduate Record Examination [GRE] Verbal and SAT Mathematical-GRE Quantitative). These findings indicate that it is the nature of the criterion, academic performance, as measured by GPA, that is changing rather than the abilities of the individuals, at least as measured by these relatively broad, aptitude tests.

A recent study by Butler and McCauley (1987) presents results that do not fit the consistent findings described above. Their annual GPAs show high, consistent correlations across all time periods, and thus do not fit the simplex model. In addition, prediction of the GPAs by aptitude test scores and high school rank do not show the characteristic decline across the four undergraduate years, but rather remain stable. The authors argue that their findings may have been due to the fact that their sample was from an institution quite different from those in the earlier work—that is, a military institution, the United States Air Force Academy, rather than civilian institutions of higher learning. Humphreys (personal communication, 4 March 1989) has characterized these findings as "puzzling" and hypothesizes that the results can be explained if cumulative grade averages, rather than independently computed grade averages, are actually used as reported in the article.

All of these studies were carried out on college students using academic performance (GPA) as the criterion. It is likely that the population of applicants for enlisted military positions differs from such a population in several important ways (e.g., type and quality of elementary and secondary education received and socioeconomic level of family) that may have an impact on performance on aptitude tests and in educational settings. On the other hand, both populations are from the same age group and are entering settings where they are expected to perform adequately or better in school settings. It seems to us, on balance, that these findings are relevant and informative for our question, but we must be cautious about the amount of weight to be placed on the findings.

We think that the major lesson to be learned from these studies is the fairly large impact on the validity coefficient exerted by the timing of measurement of the criterion. If the nature of performance on the criterion is changing and if individuals' abilities are stable, then the time at which criterion measurement is done can have a substantial effect on the observed correlations with any set of ability measures. Obviously, this could have a similarly substantial effect on attempts to calculate incremental validities. Selection of the appropriate time to measure the criterion would thus seem to be very important and should probably be driven by important organizational needs (such as the need to know if an individual has learned what is essential to perform a set of tasks) rather than by administrative convenience, available time to collect data, or other less important considerations. If criteria are measured at organizationally important points in time, then any influence of that timing on the estimation of incremental validity should be acceptable from the organization's point of view.

Additional Research on Declining Validity Coefficients

Henry and Hulin (1987) summarize and evaluate evidence pertinent to the "changing task" versus "changing person" hypotheses put forth to explain the decline in validity coefficients over time. They briefly review earlier work that they believe showed widespread declines in validity over time, including the work described just above, but also in the areas of perceptual-psychomotor performance (e.g., discriminant reaction time, tracking, rotary pursuit), student pilot performance, salary of engineers, and productivity of scientists. They then explicate the two competing hypotheses and evaluate the evidence for them. They state that the evidence to date is inconclusive,

and then present an analysis of data on professional baseball players' performance, which exhibits the simplex or superdiagonal stability matrix (a term used to denote time-related data that are simplex in form). They conclude that assumptions of stability of long-term predictive validity are unwarranted and that limited evidence is available to support the "changing subject" hypothesis.

These views are not uncontested. Barrett, Caldwell, and Alexander (1985) review and reanalyze research pertinent to three definitions of dynamic criteria: changes in group average performance over time, changes in validity over time, and changes in rank-ordering of scores on the criterion over time. They conclude that the evidence for changes in validity or changes in rank order on the criteria over time is weak and attribute the available positive evidence to methodological artifacts such as "temporal unreliability and restriction of range" (p. 53).

Austin, Humphreys, and Hulin (in press) criticize Barrett et al. for using statistical tests with low power when comparing correlations, failure to consider the entire matrix of correlations (as opposed to making only pairwise comparisons of correlations), and avoiding the issue of explaining temporal changes in correlations between criterion scores by renaming the phenomenon an artifact, "temporal unreliability," to be dismissed.

Ackerman (1989) has criticized Henry and Hulin (1987) on several grounds. He accepts the view that within-task intercorrelations fit the simplex form, but argues that this does not automatically imply a decline in validity coefficients for ability measures. He also argues that restriction of range in task performance with practice could account, at least in part, for the decline in validities. He points to several studies that show instances of increasing validities over time periods for some abilities. These findings, he argues, do not fit with the view that validities must necessarily decrease over time. He also challenges Henry and Hulin's use of job sample performance or early job/task performance as equivalent to performance on an ability test. If this assumption of equivalence is not accepted, and he argues that it should not be, then data from time-period by time-period stability matrices do not address validity coefficient data. Finally, Ackerman offers his own theory (Ackerman, 1984, 1986) as an alternative explanation for the phenomenon of decreasing validities of general intelligence kinds of measures for early (in time) task performance and increasing validities of lower-order abilities (perceptual speed and psychomotor abilities) for later (in time) task performance, which he posits as the more accurate description of the empirical results. In brief, Ackerman's theory states that task type interacts with predictor type in the determination of validity coefficients over time. Tasks that are consistent (i.e., can become automatized) will show decreased validity coefficients over time for measures of general intelligence (Ackerman [1989] cites as examples the constructs of educing relations, deriving relations, and memorizing relations.) and will show increased validities over time for lower-order ability measures. Inconsistent tasks, however, will show no attenuation over time for general intelligence measures since no automatization occurs.

Henry and Hulin (1989) respond to Ackerman's criticisms primarily by disputing his claim that their equating of ability measures with early task performance is unjustified. Both studies cite prior authors as supporting their views, and there does not appear to be compelling weight to support one view or the other. Henry and Hulin respond to Ackerman's presentation of studies that do not show decreasing validity for abilities over time with the news that they are conducting a meta-analysis of such studies, and results to date support the contention that decline in validity is ubiquitous. They also criticize Ackerman for using cross-sectional data, arguing that only longitudinal data are appropriate for examining validity changes since they include within-subject changes whereas cross-sectional data do not.

Implications for the Research Question

The decline in validity coefficients of ability measures for predicting task performance is a matter of some debate. Whether or not the phenomenon exists at all is questioned by some (Barrett, Caldwell, & Alexander, 1985), and the generality or ubiquity of the decline is questioned by others (Ackerman, 1989). Certainly, the possible explanations for the phenomenon have not been extensively investigated. Henry and Hulin (1987) have described and evaluated the "changing task" and "changing person" explanations, and present some evidence supporting the "changing person" explanation. Recall, however, the evidence from the college sample data (Angoff & Johnson, 1988; Lin & Humphreys, 1977) that seems to support the "changing task" explanation. Ackerman has presented a theory that incorporates the nature or type of task as a major part of the explanation along with the type of ability. His review and reanalysis of pertinent data is convincing (Ackerman, 1987), but his theory requires more extensive testing. One major point to be concluded from his work, however, needs emphasis:

> Given that some abilities increase in correlation with individual differences in performances over practice even as the correlations decrease with other abilities, choice of appropriate ability measures will certainly determine the ultimate global patterns of increasing, decreasing or stable correlations between intellectual abilities and task performance over practice. (p. 24)

This view is more complicated than a view that supposes a consistent decline in validities for all abilities.

Several points are relevant to estimating incremental validities. First, there is no clear consensus about the fate, over time, of validity coefficients of ability measures for predicting performance. Opinions, with accompanying supporting evidence, can be found that advocate a consistent decline in validity for virtually all ability measures, no decline for virtually all ability measures, and sometimes a decline, sometimes an increase, and sometimes stability—depending on the ability and the task. Second, it is possible that the timing of criterion measurement is likely to have differential effects on validity coefficients for ability measures. Third, consideration of the likely relationship of an ability to stage of task or job performance (i.e., early learning, late learning, or "journeyman" performance) should probably play a part in selecting predictors for a selection battery and in deciding the criteria for evaluating the "success" of a predictor.

Nature of the Criterion

Within-job multidimensionality of criteria is a complicating factor. Although predictions of job performance must ultimately be reduced to a single prediction, thus implying a single criterion of job performance, many writers conceive of job performance as multidimensional (Campbell, 1986; Dunnette, 1963; Schmidt & Kaplan, 1971). If there is more than one important, non-redundant criterion, then the choice of which criterion to measure or predict can have a decisive effect on the estimation of incremental validity for a given set of predictors. If several or all important, non-redundant criteria are measured and intended to be predicted, then the method of combining the criteria into a final composite can have a major effect on the estimation of incremental validity for a set of predictors. The dimensionality of job performance can be ignored, or assumed to be largely unidimensional, as it is when a single score such as a written test of training knowledge is used as the criterion to be predicted. If the single score does capture all or almost all of the concerns of the organization with regard to performance on a job, then this seems to be a reasonable way to

proceed. Where this is not the case, a preferable procedure would be to identify the important, non-redundant, and measurable dimensions of performance and then, if appropriate, overtly combine them for the purpose at hand (in the present case for use as a criterion in selection validation research).

Campbell (1986) presents some relevant data and results from Project A, a large-scale, selection validation research project sponsored by the Army Research Institute. In that project, five job performance dimensions were constructed from a large number of job performance measures. (Soldiers completed written job knowledge tests, hands-on performance tests, rated themselves and their peers on several rating scales, and were rated by their supervisors on several rating scales. The criterion data collection process consumed 12 hours for each soldier. The five summary dimensions generalized well across nine distinct Army Military Occupational Specialties [MOSs].) The names and definitions of the five criterion dimensions are shown in Figure 1.

1. **Core Technical Proficiency**: This performance construct represents the proficiency with which the soldier performs the tasks that are "central" to the MOS. The tasks represent the core of the job and they are the primary definers of the MOS. For example, the first-tour Armor Crewman starts and stops the tank engines; prepares the loader's stations loads and unloads the main gun; ebriosities the M60A3; engages targets with the main gun; and performs misfire procedures. This performance construct does not include the individual's willingness to perform the task or the degree to which the individual can coordinate efforts with others. It refers to how well the individual can execute the core technical task the job requires, given a willingness to do so.

2. **General Soldiering Proficiency.** In addition to the core technical content specific to an MOS, individuals in every MOS also are responsible for being able to perform a variety of general soldiering tasks—for example, determines grid coordinates on military maps; puts on, wears, and removes M17 series protective mask with hood; determines a magnetic azimuth using a compass; collects/reports information - SALUTE; and recognizes and identifies friendly and threat aircraft. Performance on this construct represents overall proficiency on these general soldiering tasks. Again, it refers to how well the individual can execute general soldiering tasks, given a willingness to do so.

3. **Effort and Leadership.** This performance construct reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient? While appropriate knowledges and skills are necessary for successful performance, this construct is meant only to reflect the individual's willingness to do the job required and to be cooperative and supportive with other soldiers.

4. **Personal Discipline.** This performance construct reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this construct show a commitment to high standards of personal conduct.

5. **Physical Fitness and Military Bearing**. This performance construct represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

Figure 1. Definition of the five job performance constructs from Project A (from Campbell, 1986).

An important point is the finding that this structure of five dimensions applied to all nine MOSs, and that the content of four of the dimensions was essentially the same for all the MOSs, while one was MOS-specific (Core Technical Proficiency). Thus, there were several important job performance dimensions that were found on all the Army jobs.

Incremental validities were computed for experimental predictors, when added to the ASVAB, for prediction of these five criteria for the nine Army MOSs. These data are shown in Table 1. Several points need to be made about this table. First, the experimental predictors were administered at the same time as the criterion data were collected, while the ASVAB scores are from the soldier's record, administered prior to entry into the Army. (The criterion and experimental predictor data were collected on soldiers with 12-24 months of service.) Second, the ASVAB's ten subtests were combined into four unit-weighted composites prior to the computation of multiple regression equations. These two factors probably operate to inflate the estimates of incremental validity over the ideal case in which the experimental tests would be administered at the same time as the ASVAB and all ten subtests would enter the equation for the ASVAB. On the other hand, the validity coefficients have been corrected for range restriction, which may operate to underestimate the true validity for the experimental predictors, since their corrections are based on their relationship to the ASVAB, which is very low for many of the experimental predictors (Campbell, 1986, p. 153). (We note that a longitudinal data collection procedure that overcomes these problems has just been completed on this project.)

Table 1

Mean Incremental Validity [a, b] for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs

	Predictor Domain					
Job Performance Construct	General Cognitive Ability (K = 4)[c]	General Cognitive Ability Plus Spatial Ability (K = 5)	General Cognitive Ability Plus Perceptual-Psychomotor Ability (K = 10)	General Cognitive Ability Plus Temperament (K = 8)	General Cognitive Ability Plus Vocational Interests (K = 10)	General Cognitive Ability Plus Job Reward Preferences (K = 7)
Core Technical Proficiency	.63	.65	.64	.63	.64	.63
General Soldiering Proficiency	.65	.68	.67	.66	.66	.66
Effort and Leadership	.31	.32	.32	.42	.35	.33
Personal Discipline	.16	.17	.17	.35	.19	.19
Physical Fitness and Military Bearing	.20	.22	.22	.41	.24	.22

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.
[b]Incremental validity refers to the increase in *R* afforded by the new predictors above and beyond the *R* for the Army's current predictor battery, the ASVAB.
[c]K is the number of predictor scores.

Given these caveats, it still seems clear that the incremental validities are different for the five criteria. The incremental validities are generally low for Core Technical and General Soldiering proficiency (.00 - .03), but are higher for the other three criteria, especially when the temperament scales (combinations of personality and biodata kinds of constructs are measured on these scales)

are added to the ASVAB. In those cases, the incremental validity ranges from .11 to .21. Thus, if only the first of these two criteria had been measured, there would be little evidence of incremental validity for the temperament measures.

Arguments can be made about the relative importance of the five criteria in this example, and about faults in the research design, but we intend the example simply to provide an empirical illustration that the choice of job performance dimensions for criteria can have a large impact on estimates of incremental validity.

Statistical and Design Considerations

General Considerations About Validation Research Design

Barrett, Phillips, and Alexander (1981) review the history of the distinctions made between concurrent and predictive validity designs along with the criticisms often made of the concurrent designs. They question the supposed superiority of predictive validity research designs over concurrent validity research designs on logical and empirical grounds. They point out that empirical results indicate that, for cognitive tests at least, validity coefficients are nearly equal for the two designs. Thus, they say, "We must conclude that either these potential distortions [of the two types of research designs] are trivial in their effects and act equally in the two designs, or that individual effects act unequally but counterbalance themselves across the two designs" (p. 5).

Guion and Cranny (1982) respond by describing the two designs in more detail, in effect expanding the number of designs to five, and conclude that the different validity designs are not equivalent on either conceptual or practical bases.

Sussmann and Robertson (1986) provide a more complete presentation of validation study designs. They describe and evaluate 11 criterion-related validation designs in terms of external, internal, construct, and statistical validity dimensions. Of their 11 designs, 4 are unfeasible for use in military selection since they call for random selection among applicants. With regard to the hypotheses of Temporal Contiguity (TC) and Better Ability Estimate (BAE), their discussion of time and its effect on research design is most relevant. They point out that the concurrent class of designs are most appropriate for use when time-related factors are regarded as serious alternative explanations or confounds in validation research, because no (or very little) time has elapsed between predictor and criterion administrations. However, as they also point out, when the predictor construct is viewed as an aptitude (i.e., it is viewed as predicting future job performance after a training program), then a predictive validation design is the only logical choice. Both of these factors, time-related changes viewed as alternative explanations and predictor constructs viewed as aptitudes, are appropriate for our problem. Thus, another of their conclusions seems appropriate: "Finally, to the degree that different designs yield different advantages, researchers should consider validation research programs that would use more than one validation study, each with a different design, to enhance the validity of the research" (p. 467).

It appears that more than one study will be necessary to attempt to disentangle and evaluate the TC and BAE hypotheses in the Navy selection setting. We outline our recommendations for those studies in the conclusions and recommendations section.

Statistical Considerations

Wolfe (1988) has described issues and problems confronted in estimating incremental validities for experimental predictors over and above the ASVAB. We have very little to add to his thorough discussion. We briefly address some power considerations given the likelihood of different levels of incremental validity, on the order of those observed on the Army's Project A (Campbell, 1986).

Wolfe (1988) conducted a power analysis concerning the F-test for incremental validity, and he presents estimates of the sample sizes needed for a 90% chance of detecting incremental validity of new cognitive and perceptual tests over the ASVAB. These estimates are based on a multiple R for the ASVAB alone of .59 and a multiple R after adding the new predictors of .61 (an incremental validity of .02). The results indicate that sample sizes of about 500 will be needed. After taking into account an estimated 25% attrition rate for the sample and random responding on the experimental tests by 25% of the sample, Wolfe estimates that initial sample sizes of 1000 will be needed.

However, as discussed earlier, research by the Army (Campbell, 1986) has found incremental validities for four temperament composites ranging from .11 to .20 for those job performance criteria that are not well predicted by the ASVAB (Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing). Sample sizes needed to detect these incremental validities will be smaller. For example, for the Effort and Leadership criterion, the validity of the four ASVAB composites is .31, and the validity of these four ASVAB composites plus four experimental temperament composites is .42. The sample size needed to detect this level of incremental validity using a significance level of .05 is only 173; using a significance level of .01, a sample size of 228 is needed. When these are corrected for attrition and random responding (assuming the estimates used by Wolfe), estimates for the initial sample sizes are 308 and 406 respectively.

For Physical Fitness and Military Bearing, the estimated sample sizes needed to detect incremental validity are even smaller. The validity of the four ASVAB composites for this criterion is .20; after adding the four temperament composites, the validity is .41. A power analysis indicates that a sample of 115 would be needed to detect this increase at the .05 level of significance and 150 would be needed at the .01 level. Correcting for attrition and random responding yields estimated initiated initial sample sizes of 205 and 267 respectively.

These computations illustrate the point that increases in sample sizes beyond those already contemplated by the Navy are not necessary in order to have acceptable levels of power for detecting incremental validities of non-cognitive kinds of predictors for criteria important to military jobs. Indeed, if the study were to focus on predictors and criteria like those described in Project A, smaller sample sizes would be sufficient. Increased effort and expense would be necessary, however, to determine the appropriateness of similar criteria and predictors for the Navy setting and to develop and collect the appropriate criterion measures.

The statistical considerations inherent in evaluating the viability of the TC and BAE hypotheses are not the same as those in estimating incremental validity, concerning primarily the comparison of multiple correlation coefficients for different groups. We describe those considerations in concert with the research design considerations in the conclusions and recommendations section.

Conclusions and Recommendations

Summary of Conclusions from Literature Review

The review of evidence concerning the stability of scores for young adults on measures of individual differences like those that will be used in military selection settings leads to the following conclusions:

- **Trait stability.** Scores on cognitive ability measures like the "power" scales on the ASVAB (i.e., all but the Numerical Operations and Coding Speed tests) are extremely stable, but not uniformly so, with true-score correlations over a one-year period on the order of .95-.98.
- **Perceptual and psychomotor abilities.** Perceptual and psychomotor measures provide scores that appear to be not quite as stable as cognitive abilities, with crude estimates of true-score correlations over a one-year period of .75-.90 (based on GATB data only; U.S. Department of Labor, 1970). There are larger differences in this class of measures with some computer-administered measures that use timing of motor responses and reaction times as scores showing two-week, retest coefficients in the .30s to .60s.
- **Biographical questionnaires.** There is much less evidence available for scores on biographical data measures, but what there is seems to indicate that they are very stable, with true-score correlations over a one-year period of about .90 (this estimate is based only on two studies of Owens' work with college samples).
- **Personality traits.** As we noted earlier, personality scales are extremely diverse with respect to constructs measured, names applied to the same constructs, and scales devoted to the same construct. Also, immediate or short term, test-retest reliabilities for many of these instruments are lower relative to other types of measures (e.g., about .50 for some measures). Others are higher, but most are no higher than the .70s and .80s. When these reliability limitations are taken into account, the observed stability coefficients over several years of .50 and .60 are somewhat more impressive, but still much lower than for other areas of measurement.
- **Vocational interests.** Vocational interest scale scores appear to be highly stable for persons over age 22 (three-year stability coefficients in the low .80s), but only moderately stable for persons between the ages of 17 to 21 (three-year stabilities in the .60s).

These conclusions indicate that the probability that true score changes could account for changes in validity coefficients is clearly related to the construct under consideration. For the cognitive ability constructs measured by the ASVAB, the probability seems extremely low that true score changes could account for any substantial validity changes over periods of six months to three years, which include the probable periods of research appropriate for first-term enlisted military personnel. (True score changes on cognitive abilities different than those on the ASVAB could be more or less probable.) The probable amount of true score change in other domains is higher than for cognitive ability and, therefore, is a stronger candidate for explaining changes in validity coefficients for those measures.

Our review of the evidence concerning the impact of intervening events (between test administrations) on test scores and their relationships with other variables leads us to conclude:

- **Maturation.** Maturation probably has little effect on the stability of individual difference measures for the young adult population.
- **Life experiences.** Exposure to vocational or academic training after entry into the Navy is unlikely to have a noticeable effect on the stability of ASVAB-like measures.
- **Test sophistication.** Changes in test sophistication or test-wiseness are also viewed as likely to have little impact on stability or validity coefficients for the military applicant population.
- **Practice and coaching.** Traditional practice and coaching effects have small to moderate positive effects on scores, though the evidence is mixed, and such effects may be moderated by length and type of coaching as well as by the nature of the test. Even where such effects on test scores do exist, most researchers seem to have assumed they probably have little impact on the validity of test scores, especially if coaching or practice is either absent or present for all or almost all test-takers. Embretson (1987) has shown that, for at least one type of test (spatial ability), a dynamic testing procedure can lead to improved test reliability and validity for an external criterion. This finding is in contrast to the view that coaching-like interventions should reduce validity (Cole, 1982).
- **Motivation and other temporary states.** Changes in motivation and other temporary states from one testing administration to another, though certain to occur, are not a concern for validity unless there are systematic changes for a substantial number of test-takers. It is difficult to generalize about the possibility of such occurrences. Persons can alter their scores on tests and inventories when instructed to do so, but the extent to which they do so in "real-life" and "experimental" settings is still largely unknown, with some evidence that such reactions are less common than one might suppose (Peterson, 1987).

In line with this summary, we think that maturation and an increase in general test sophistication are unlikely to have any effect on ASVAB test scores. The probable effects of practice, coaching, and dynamic testing (for the military applicant population) are not clear at all and may be specific to type of test and intervention. Motivational changes large enough and systematic enough to affect validities seem unlikely, though not impossible, and, again, may be different for different types of measures.

The evidence summarized above does not strongly support either the TC or the BAE hypothesis. On balance, we think the evidence is slightly in favor of the BAE explanation for any changes in ASVAB validities due to a second administration of the ASVAB, especially if the intervening time interval is too short to allow a high likelihood of true score changes (i.e., less than one year or so).

Our review of the literature concerning changes over time in the validity of ability measures for predicting performance leads us to conclude that there is no simple relationship between the level of obtained validity coefficients and the amount of time between the administration of the predictor measures and the collection of the criterion data. Many researchers argue that aptitude tests consistently show decreasing validities over time for predicting a variety of criteria (e.g., job performance, college grades, etc.), but others argue that this conclusion is not justified by the available evidence. A few laboratory studies have actually found that while the validity of some abilities for predicting performance decreases over time, the validity of other abilities (e.g., lower order abilities such as perceptual speed and psychomotor abilities) increases.

Even in those cases where the size of validity coefficients clearly decreases over time, available research does not explain this decrease. It seems likely that the reason for decreases in obtained validities over time differs between studies. Evidence from college student sample data supports the "changing task" explanation, but other evidence seems to support the "changing person" explanation. In addition, since job performance can be seen as multidimensional, abilities that are strongly related to one important aspect of job performance may be less related to others, and the stability of these validity coefficients over time is likely to differ for the various aspects of job performance as well. Thus, the job performance measure(s) chosen, the timing of the measurement of job performance, and the predictor measure(s) chosen will interact in determining the stability of obtained validity coefficients over time. All of these factors must be taken into account in determining whether or not decreases in obtained validity coefficients are to be expected over time.

If all validity coefficients of interest were stable over time, there would be no problem in estimating the incremental validity of experimental predictors administered at a different time than the operational predictors. However, based on the evidence summarized above, even if the validity of the operational predictors was found to be stable over time for a given measure of job performance, this would not guarantee that the validity of a set of new experimental predictors would be stable over time or that the validity of the operational predictors for a different measure of job performance would be stable over time.

Different validation designs also yield different information concerning the relationship between "predictor" measures and job performance criteria. It appears likely that several studies would be necessary to understand the relationship between the predictor measures and important job performance criteria.

Statistical considerations in estimating incremental validities have been thoroughly summarized by Wolfe (1988). The present review makes the additional point that in order to detect the incremental validities of some alternative predictors for certain aspects of job performance (e.g., the validity of temperament and biodata scores for Personal Discipline in the Army; Campbell, 1986), substantially lower sample sizes may be needed than those needed to detect the incremental validities of additional cognitive and perceptual measures in predicting training knowledge.

Research Design Considerations and Recommendations

Importance of Criterion Selection for Validation Research

We have pointed out that the choice of criterion measure and time of measurement of the criterion can affect the interpretation of validity coefficients. The straightforward solution to this problem is to use those criteria that are of the highest organizational importance, however defined, and measure them at the time that the criterion information is most useful for organizational decision making.

Successful completion of military training schools seems to fit that specification and is routinely measured and utilized by the armed services. Many, if not most, training programs are completed within 6 to 12 months of entry into the Navy. This period of elapsed time does not seem long enough to allow true score changes to affect validity coefficients for most of the ASVAB measures, as we have noted. Nevertheless, this simply results in a lower prior probability for support of the TC hypothesis for explaining changes in ASVAB validities for this criterion and does not argue against the use of training success in such research.

Successful job performance at the journeyman level in Navy ratings is also a highly important organizational outcome, but the nature and definition of this criterion is likely to be more difficult to identify. Also, the point in time at which success can or should be expected is probably much less clear than it is for training success. It will always, of course, be at a point later in time than training success and therefore would afford more opportunity for true score changes to affect ASVAB validities. Once again, this is not an argument for or against the use of journeyman job performance in research, but results in a higher prior probability for support of the TC hypothesis. At any rate, it seems to us that the choice of criteria in validation research of any kind should be made for the reasons we have noted—which are independent of the effects of the criterion on evaluation of the validity of candidate predictors.

Another implication of the differential effects of criteria on validity estimates is that all testees should be measured with the same criterion. This would seem to limit analyses of research data to samples from the same ratings, but only if the criterion differs across rating. The Project A research (Campbell, 1986) indicates that some criteria may be both important and conceptually similar across jobs. However, scaling differences can severely complicate analyses on pooled samples even when the same measurement instruments are used across subjects. Thus, it seems that, practically speaking, within-rating subjects will usually be necessary.

Research Designs for Temporal Contiguity (TC) versus Better Ability Estimate (BAE) Hypotheses

Figure 2 shows a research design for investigating the viability of the TC and BAE hypotheses. In this design, one group of Navy recruits would be readministered the ASVAB close in time to both entry into the Navy and the first administration of the ASVAB, no more than 2 to 4 weeks after the first administration. Another group of Navy recruits would be readministered the ASVAB several months to two years after entry into the Navy and the first administration of the ASVAB, but the second administration of the ASVAB would be no more than 2 to 4 weeks from the collection of criterion data for the group. For both groups, the criterion data would be collected at the same point in their career—perhaps at the end of training or at a point in their career when job performance is expected to be at the journeyman level—but on an occasion separate from the second administration of the ASVAB. The Navy recruits would be randomly assigned to either of the two groups from the same population, probably defined by rating. For example, one half of the recruits assigned to Rating A would go into Group 1 and be given the early readministration while the other half would go into Group 2 and be given the later readministration of the ASVAB, but the total sample would be administered the same criterion measure.

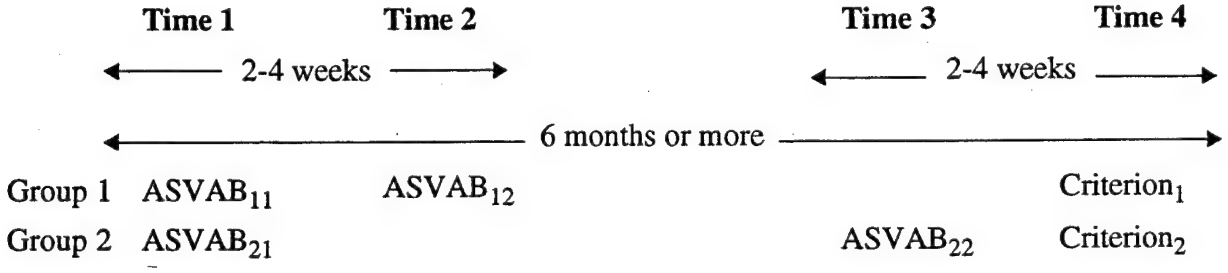

Figure 2. Possible research design for investigation of Temporal Contiguity (TC) and Better Ability Estimate (BAE) hypotheses.

The second administration of the ASVAB will provide higher validity for predicting the criterion than will the first administration for both Groups 1 and 2, according to the BAE hypothesis, while the TC hypothesis holds that validities will increase only for Group 2 (or that Group 2 validities will show an increment beyond that shown for Group 1). Only the BAE hypothesis can account for improvements in validity for Group 1 since not enough time has elapsed between ASVAB administrations for true score changes to occur, thus any improvements in validity must be attributed to reduction in error of measurement. Group 2 represents the TC hypothesis, where a larger amount of time has elapsed between the two administrations so that true score changes, if any, can operate to improve the validity.

Figure 3 shows the matrix of predictor scores (X matrix) represented by the research design. These scores are used to predict Y, the criterion scores, which can be represented:

$$Y = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1N} \\ \hline y_{21} \\ y_{22} \\ \vdots \\ y_{2N} \end{bmatrix}$$

$$X = \underset{(2N \times 32)}{} \left[\begin{array}{lll} 1\,0\,X_{1111} \ldots X_{111m} & X_{1211} \ldots X_{121m} & 0 \ldots 0 \\ 1\,0\,X_{1121} \ldots X_{112m} & X_{1221} \ldots X_{122m} & 0 \ldots 0 \\ 1\,0\,X_{1131} \ldots X_{113m} & X_{1231} \ldots X_{123m} & 0 \ldots 0 \\ \vdots & \vdots & \vdots \\ 1\,0\,X_{11N1} \ldots X_{11Nm} & X_{12N1} \ldots X_{12Nm} & 0 \ldots 0 \\ \hline 0\,1\,X_{2111} \ldots X_{211m} & 0 \ldots 0 & X_{2211} \ldots X_{221m} \\ 0\,1\,X_{2121} \ldots X_{212m} & 0 \ldots 0 & X_{2221} \ldots X_{222m} \\ 0\,1\,X_{2131} \ldots X_{213m} & 0 \ldots 0 & X_{2231} \ldots X_{223m} \\ \vdots & \vdots & \vdots \\ 0\,1\,X_{21N1} \ldots X_{21Nm} & 0 \ldots 0 & X_{22N1} \ldots X_{22Nm} \end{array}\right]$$

Note. First subscript denotes group membership; second subscript denotes first or second administration; third subscript denotes subject within group; fourth subscript denotes ASVAB subtest, m = 10.

Figure 3. Representation of matrix of predictor data.

The linear prediction model is Y = XB + E. The column vector of 32 parameter estimates is shown below.

$$B = \begin{bmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_m \\ c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_m \\ d_1 \\ d_2 \\ d3 \\ \vdots \\ d_m \end{bmatrix}$$

Thus, if all the information in the predictor matrix is used to predict Y, the criterion scores, then 32 parameters are estimated—one for each of the group intercepts (a_1 and a_2), ten ASVAB scores for the combined Groups 1 and 2 (b_1 to b_m, with m = 10), the second administration for Group 1 (c_1 to c_{10}), and the second administration for Group 2 (d_1 to d_{10}).

We are interested in the degree to which validity (R^2) decreases by the use of various constrained or restricted models, representing hypotheses of interest. There are three hypotheses of interest. The first is the null hypothesis that a second administration results in no increase in validity and will be labeled the General Second Administration hypothesis. If this hypothesis is rejected, we would conclude that an increase in validity did occur, and we can make additional tests concerning the BAE and TC hypotheses. In the terms of our model, this hypothesis states that $c_1 \ldots c_m$ and $d_1 \ldots d_m$ are constrained to zero. The rank of this constrained model is 12 (32-20). The statistical test used is the F-test for comparing full and restricted models:

$$F_{k,\, n-p-k-1} = [(R^2_{p+k} - R^2_p)/k]/[(1-R^2_{p+k})/(n-p-k-1),$$

where

R_p = the multiple correlation of the p predictors with the criterion

R_{p+k} = the multiple correlation of the p+k predictors with the criterion

n = sample size.

For the General Second Administration hypothesis, p = 12 and k = 20, as we just described, and n = 2N, where N = size of Group 1 or Group 2 (equal group sizes).

The second hypothesis of interest states that the later administration (Group 2 administration) adds no additional validity beyond that added by the earlier administration and will be labeled the Temporal Contiguity (TC) hypothesis. This is the most important test. If this hypothesis is rejected, then the TC hypothesis is credible, and true score changes would appear to affect validity of the ASVAB over and above the effect due to a retest on the ASVAB. In terms of our model, we would estimate a new set of B weights constrained so that $c_1 = d_1 = f_1$, $c_2 = d_2 = f_2$, . . . , $c_m = d_m = f_m$, where f_i is the new weight corresponding to the i^{th} subtest. Thus, the restricted model constrains the weights for the earlier and later administrations to be equal and has rank 22; $p = 22$ and $k = 10$ for the F-test.

If the Navy is interested in increments in validity to the ASVAB, from whatever source, on the order of .02 (Wolfe, 1988), then we should employ sufficient N to reliably detect an increment due to Temporal Contiguity as small as .01. If we cannot reject the null hypothesis with a test of high power for detecting this level of difference, then we can be assured that no more than half of statistically significant incremental validities of .02 from other measures could be due to true score changes on constructs measured by the ASVAB. To compute the required Ns, then, we have assumed a full model R of .60 (a reasonable assumption of the validity of the ASVAB for organizationally important criteria; Wolfe, 1988), a desired power of .90 for detecting differences as small as .01, and alpha levels of .01 and .05. Given the models described above, the resulting required Ns are shown below (these estimates are based on procedures from Cohen and Cohen, 1975, p. 145). Note also that the "Required N" refers to the total sample size (assuming a balanced design, equal Ns for groups 1 and 2).

For the General Second Administration hypothesis (k = 20):

alpha =	.01	.05
Required N =	1914	1488
Estimated Initial N = (if 25% attrition)	2552	1984

For the Temporal Contiguity hypothesis test (k = 10):

alpha	.01	.05
Required N =	1528	1173
Estimated Initial N = (if 25% attrition)	2037	1567

If the General Second Administration test is not significant, then neither TC nor BAE are operating, and we can assume that second administrations of the ASVAB will have no effect on estimates of validity for the ASVAB. If the first test is significant, but the second test is not significant, then we can assume that a second administration of the ASVAB does affect estimates of validity, but that true score changes do not account for this effect (i.e., the BAE hypothesis has won the day). If the first and second tests are significant, then we can assume that a second administration of the ASVAB does affect estimates of validity *and* that a significant part (at least .01) is due to true score changes.

However, in this last case, we do not know whether or not BAE is operating in addition to TC. In order to conclusively determine whether or not BAE is operating in this case, one additional test

is needed that involves only Group 1. This is a test of the null hypothesis that immediate readministration does not result in an increase in validity. This test, however, can only be done using Group 1, since time of administration and readministration are confounded in Group 2. The full model in this case contains the X matrix for Group 1 only, and a_1, $b_1 \ldots b_m$, $c_1 \ldots c_m$ in the B vector (there is no set of d weights corresponding to the second administration to the second group). This full model rank = 21. The constrained model sets $b_1 = c_1 = g_1$, $b_2 = c_2 = g_2$, $\ldots$ $b_m = c_m = g_m$, and has rank = 11, thus p = 11 and k = 10, for the F-test. (Note also that this test is not independent of the other two tests). If this test is done, Group 1 will need to be larger in order to maintain the same level of power. This increase in size is independent of the desired sample size of Group 2, except that we wish to maintain a balanced design for the hypothesis tests, thus requiring at least one half of the necessary sample size for those tests to be in Group 2. Taking all of this into account, the following sample sizes are required in order to complete all three hypothesis tests (i.e., the General Second Administration test, the Temporal Contiguity test, and the Better Ability Estimate test):

alpha	.01		.05	
	Group 1	Group 2	Group 1	Group 2
Required N =	1517	957	1162	744
Estimated Initial N = (if 25% attrition)	2023	1276	1550	992

Two valid criticisms can be made of this design. First, including a second administration in the model will always lead to higher validity merely because the two administrations together are essentially a longer, more reliable test. Second, the sample has been explicitly selected based on the first administration of the ASVAB. The first criticism, while true, does not affect the ability of the design to address the central research questions. It only applies to the test of the General Second Administration Hypothesis. The test of the Temporal Contiguity hypothesis is not affected, since both the full and the restricted model contain two administrations. The test concerning whether an immediate second administration will increase validity will be similarly unaffected. Thus, the research design does not confound the second administration problem with questions concerning whether BAE is operating or TC is operating. The only question that is affected is whether or not there is an effect of a second administration at all.

The second criticism, concerning explicit selection, however, is confounded with a question of interest. Since the sample is selected based on their first administration of the ASVAB, we can expect the second administration to have larger variances and, consequently, slightly higher validities. If the research shows that BAE is true, there is no way of knowing to what extent this is due to the fact that the second administration is truly a better ability estimate or to what extent this is merely an artifact of explicit selection on the first administration. The only sure way to get around this would be to eliminate no one based on the first administration of the ASVAB, and this is not practically feasible. The effect of explicit selection on the obtained validities could, to some extent, be assessed by correcting the obtained correlation matrices for range restriction. In other words, validities obtained from the first and second administrations would be corrected so that both were comparable to the total population. These validities could then be compared to the uncorrected validities in order to determine whether conclusions concerning TC and BAE would be changed. Although not widely known, there are methods for testing the significance of corrected (e.g., for attenuation or range restriction) correlation coefficients (Bobko, 1983; Bobko & Rieck, 1980). However, APA guidelines generally counsel against using this correction (APA, 1974). Another approach would be to eliminate from the data set those

examinees who obtained "failing" scores in the second administration of the ASVAB; that is, they would not have been selected into their job had they obtained that score on the first administration. Aside from throwing away data, it could be argued that the people excluded using this method are different in some systematic way from the rest of the population and the resulting sample is less representative than if they had been included. Thus, there seems to be no really satisfactory answer to this criticism. The approaches employing range restriction correction and deletion of second administration "failures" are not completely satisfactory, and not selecting on the basis of ASVAB scores is not feasible.

In addition, sample sizes required for this design are large, and it is likely that criterion considerations (see above) will restrict sampling within ratings. It could take a full year or more to obtain these sample sizes within even the larger ratings. There will also be some limits on generalizability of the findings with regard to the length of the time interval used to allow true score changes (i.e., the Group 2 interval between ASVAB administrations) and the type or nature of criteria used in the study and its relationship to ASVAB and likely additional predictors. In addition, the design does not address the effects of interventions like coaching or dynamic testing, although we do think it addresses the most pertinent "intervention" (i.e., simple readministration of the ASVAB). Of course, the described research has nothing at all to say about the likelihood of the TC hypothesis for any predictors beyond the ASVAB.

These limits could perhaps be addressed by additional studies or an expanded version of the design we have described. For example, one study could use end-of-training criteria for a fairly long school (one year or so) while another could use end-of-training criteria for a fairly short school (say less than three months). Another study could use on-the-job performance criteria administered just at the end of the usual first-term, enlistment period. This series of studies might shed additional information on the degree to which the TC and BAE hypotheses hold up across various time intervals and types of important criteria, but they create additional practical (the same Ns are needed for these studies as outlined for our example) and design problems (length of school is probably confounded with difficulty and complexity of subject matter/job tasks) of their own.

Additional interventions beyond simple readministration (such as dynamic testing procedures) could be studied by using them in place of, or in addition to, the simple readministration. A third experimental group could be added to the present design to accomplish this, or a separate study using such procedures instead of simple readministration. Either method would require larger Ns. The payoff would be the demonstration that more intensive interventions might result in increases in validity so that no further increases could be accounted for by true score changes over time periods of concern. The practical implication of such a finding would be the early use of these interventions on samples that would be included in longitudinal validation studies, in order to obviate the readministration of the ASVAB at the later point in time when the criterion data are collected. Given the present state of knowledge about dynamic testing in adult populations for variables such as those found on the ASVAB, we would be reluctant to recommend this type of research without first completing the study we have outlined.

Christal (1989) describes a research design that suggests another approach to the research question. In this study, the ASVAB was readministered to a random sample of 4077 Airmen on the sixth day of their Basic Military Training (Time 2). These Airmen had, of course, already taken the ASVAB before entering the military (Time 1). One of the ASVAB subtests administered at Time 2 was designated the "criterion" measure; and another related subtest, the "experimental predictor." Using multiple regression, scores on this criterion were predicted using the remaining eight ASVAB subtests. The experimental predictor was then added to the regression and the

incremental validity for the experimental test was computed. This was repeated using each of the Time 2 ASVAB subtests (except Coding Speed and Numerical Operations) as the criterion. The incremental validity of the experimental predictor was examined for three different models that varied the timing of the eight ASVAB subtests, the experimental tests, and criterion measures:

Model 1: Criterion and Experimental Test administered at Time 2; ASVAB administered at Time 1.

Model 2: Criterion administered at Time 2; Experimental Test and ASVAB administered at Time 1.

Model 3: Criterion, Experimental Test, and ASVAB administered at Time 2.

For each model, the incremental validity of the experimental test over the remaining ASVAB subtests was computed for each of the eight criteria. A comparison between the average incremental validity obtained using Model 1 and that obtained using Model 2 gives an estimate of the amount of inflation in the incremental validity estimate obtained when operational ASVAB scores are used to compute the incremental validity of an experimental predictor that is administered in the same session as the criterion measure. The average estimated variance contribution of the experimental predictor in Model 1 is 12.8% while in Model 2 it is only 10.4%. Results for Model 3, where the ASVAB and experimental test are administered concurrently with the criterion are very similar to those for Model 2, which Christal interprets to mean that readministration of the ASVAB will result in reasonably accurate estimates of incremental validity. This does not, however, take into account the BAE hypothesis.

Christal's design does not directly address our research question, since validity differences due to change over time are confounded with the fact that the predictors from Time 2 and the criterion measures were administered as part of the same testing session. The fact that the validities are higher for Time 2 could be partly due to correlated error (e.g., motivation). Lloyd Humphreys (personal communication, 4 March 1989) suggested a modification of Christal's design that more directly addresses our research question. Humphreys suggested that two additional comparisons could be made using Christal's data. Rather than using only subtest scores at Time 2 as the criterion measures, subtest scores at Time 1 would be used as criterion measures as well. Two comparisons could then be made. One comparison would be the difference between the validity of the subtests scores at Time 2 for predicting criterion scores at Time 1 and the validity of subtest scores at Time 1 for predicting criterion scores at Time 2. The second comparison would be the difference between the validity of the subtest scores at Time 1 for predicting criterion scores at Time 1 and the validity of subtest scores at Time 2 for predicting criterion scores at Time 2. Finally, this line of reasoning suggests that the Time 2 scores would be more internally consistent than the Time 1 scores.

If BAE is false, there is no reason to expect the validity coefficients to differ in either comparison. However, if the BAE hypothesis is true, we would expect the ASVAB subtests administered at Time 2 to be better ability estimates. Test scores from the second administration could be seen as containing a larger proportion of true score variance. To the extent that true scores on the underlying abilities are correlated, we would expect these better ability estimates to be more highly correlated with other tests. When eight tests administered at Time 2 are used to predict a single test administered at Time 1, we would, on the average, expect higher validity than when using eight tests administered at Time 1 to predict one test administered at Time 2, since in the former case all eight predictors are better ability estimates and in the latter case only the single criterion measure is a better ability estimate. Similarly, we would expect validity of scores at

Time 2 for predicting scores at Time 2 to be higher than scores at Time 1 for predicting scores at Time 1. Following this line of reasoning, we would expect subtests administered at Time 2 to have higher intercorrelations than those administered at Time 1, and this is in fact what Christal found. The average correlation of a subtest with all other subtests is higher at Time 2 for each of the ten subtests: the overall average across all ten was .318 at Time 1 and .351 at Time 2.

Even with these additions, the Christal design can only test the BAE hypothesis. If results showed that the BAE hypothesis is not true (i.e., the validities are equal), then this research design answers our question. If, however, the results show that the BAE hypothesis is true, we cannot rule out the possibility that the TC hypothesis is also true. In addition, the design would not tell us how much of the difference in validity for an external criterion is due to TC and how much is due to BAE. Essentially, this design only tells us whether or not BAE is, to some extent, true. It is also difficult to determine the power of this design and the sample sizes needed, since it will depend on the size of the true score correlations between ASVAB subtests.

Most of these research designs would require a great deal of effort to complete. The Christal design would not require as much effort, but it does not definitively answer the research question. Given all of this, it might be most prudent to straightforwardly design the incremental validity research to avoid confounding with TC or BAE hypotheses. Figure 4 shows an "optimal" design for accomplishing this feat.

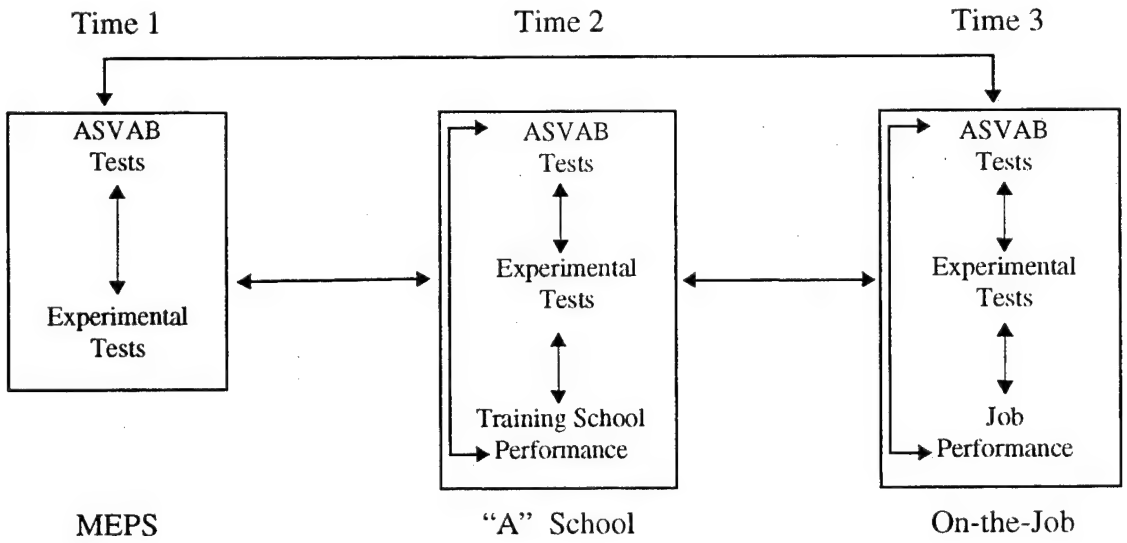

Figure 4. Optimal incremental validation design.

Although Figure 4 shows the administration of all candidate predictors at three points in time, this is not necessary if we are concerned only with predictive validity. In that case, we need only administer all the predictors at the same time (or nearly so) to all sample members, presumably at the time that the information would be used for organizational decision making. Provided that none of the predictors are readministered, this would control for either the BAE or TC hypothesis with regard to their differential effect on ASVAB and experimental validities. The additional administrations of the predictors would shed additional light on the nature of score changes in the predictors, and might prove useful if concurrent validities were organizationally useful (for example, if the Navy needed to quickly select the 100 best "gunners," and a short battery existed that had been demonstrated to have a high concurrent validity for an extended set of criterion measures).

References

Ackerman, P. L. (1984). *A theoretical and empirical investigation of individual difference in learning: A synthesis of cognitive ability and information processing perspectives.* Unpublished doctoral dissertation, University of Illinois, Urbana.

Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence, 10,* 101-139.

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3-27.

Ackerman, P. L. (1989). Within-task intercorrelations of skilled performance: Implications for predicting individual differences? (A comment on Henry & Hulin, 1987). *Journal of Applied Psychology, 74,* 360-364.

Alvares, K. M., & Hulin, C. L. (1973). An experimental evaluation of a temporal decay in the prediction of performance. *Organizational Behavior and Human Performance, 9,* 169-185.

American Psychological Association, American Educational Research Association, and National Council on Management in Education. (1974). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: MacMillan Publishing Co., Inc.

Angoff, W. H. (1988). The nature-nurture debate, aptitudes, and group differences. *American Psychologist, 43*(9), 713-720.

Angoff, W. H., & Johnson, E. G. (1988). *A study of the differential impact of curriculum on aptitude test scores* (Research Report RR-88-46). Princedom, NJ: Educational Testing Service.

Austin, J. T., Humphreys, L. G., & Hulin, C. L. (in press). *Another view of dynamic criteria: A critical reanalysis of Barnett, Caldwell, and Alexander.*

Baggaley, A. R. (1974). The stability of interest variables and items during adolescence. *Journal of Multivariate Experimental Clinical Research, 1,* 38-45.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology, 38,* 41-56.

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 66,* 1-6.

Bayley, N., & Oden, M. H. (1955). The maintenance of intellectual ability in gifted adults. *Journal of Gerontology, 10,* 91-107.

Bobko, P. (1983). An analysis of correlation corrected for attenuation and range restriction. *Journal of Applied Psychology, 68,* 584-589.

Bobko, P., & Rieck, A. (1980). Large sample estimates for standard errors of functions of correlation coefficients. *Applied Psychological Measurement, 4,* 385-398.

Butler, R. P., & McCauley, C. (1987). Extraordinary stability and ordinary predictability of academic success at the United States Military Academy. *Journal of Educational Psychology, 79*, 83-86.

Campbell, D. P. (1971). *Handbook for the Strong Vocational Interest Blank.* Stanford, CA: Stanford University Press.

Campbell, D. P., & Hansen, J. C. (1981). *Manual for the SVIB-SCII Strong-Campbell Interest Inventory.* Stanford, CA: Stanford University Press.

Campbell, J. P. (Ed.). (1986). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1986 fiscal year* (ARI Technical Report 813101). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Christal, R. E. (1989). *Estimating the contribution of experimental tests to the Armed Forces Vocational Aptitude Battery* (AFHRL Technical Paper 89-30). Brooks Air Force Base, TX: Armstrong Human Resources Laboratory.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, N. (1982). The implications of coaching for ability testing. In A. K. Wigdor, & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies (Part II: Documentation section)* (pp. 389- 414). Washington DC: National Academy Press.

Costa, P. T., & McCrae, R. R. (1978). Still stable after all these years: Personality as a key to some issues in adulthood and old age. In P. B. Baltes, and O. G. Brim, Jr. (Eds.), *Life-span development and behavior, 3*, 65-102. New York: Academic Press.

Costa, P. T. Jr., McCrae, R. R., & Arenberg, D. (1980). Enduring dispositions in adult males. *Journal of Personality and Social Psychology, 38*(5), 793-800.

Counselor's manual for the Armed Services Vocational Aptitude Battery: Form 14. (1984). Department of Defense.

Cronbach, L. J. (1984). *Essentials of psychological testing.* New York: Harper & Row.

Dann, J. E., & Abrahams, N. M. (1973). *Occupational scales of the Navy Vocational Interest Inventory: II: Reliability* (TR-74-5). San Diego, CA: Navy Personnel Research and Development Center.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53,* 1-15.

Droege, R. C. (1966a). Effects of practice on aptitude scores. *Journal of Applied Psychology, 50,* 306-310.

Droege, R. C. (1966b). GATB longitudinal maturation study. *Personnel and Guidance Journal, 44*, 919-930.

Dunnette, M. D. (1963). A modified model for test validation and selection research. *Journal of Applied Psychology, 47,* 317-323.

Eichorn, D. H., Hunt, J. V., & Honzik, M. P. (1981). Experience, personality, and IQ: Adolescence to middle age. In D. H. Eichorn, J. A. Clausen, N. Haan, M. P. Honzik, & P. H. Mussen (Eds.), *Present and past in middle life* (pp. 89-116). New York: Academic Press.

Embretson, S. E. (1987). Improving the measurement of spatial aptitude by dynamic testing. *Intelligence, 11,* 333-358.

French. J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement, 25,* 9-28.

Gough, H. G. (1975). *Manual for the California Psychological Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Guilford, J. S., Zimmerman, W. S., & Guilford, J. P. (1976). *The Guilford-Zimmerman Temperament Survey handbook: Twenty-five years of research and application.* San Diego, CA: Knapp.

Guion, R. M., & Cranny, C. J. (1982). A note on concurrent and predictive designs: A critical reanalysis. *Journal of Applied Psychology, 67,* 239-244.

Hansen, J. C., & Stocco, J. L. (1980). Stability of vocational interests of adolescents and young adults. *Measurement and Evaluation in Guidance, 13,* 171-176.

Hansen, J. C., & Swanson, J. L. (1983). Stability of interests and the predictive and concurrent validity of the 1981 Strong-Campbell Interest Inventory for college majors. *Journal of Counseling Psychology, 30,* 194-201.

Henry, R. A., & Hulin, C. L. (1987). Stability of skilled performance across time: Some generalizations and limitations on utilities. *Journal of Applied Psychology, 72,* 457-462.

Henry, R. A., & Hulin, C. L. (1989). Changing validities: Ability-performance relations and utilities. *Journal of Applied Psychology, 74,* 365-367.

Honzik, M. P., MacFarlane, J. W., & Allen, L. (1948). The stability of mental test performance between two and eighteen years. *Journal of Experimental Educational, 17,* 309-334.

Humphreys, L. G. (1960). Investigations of the simplex. *Psychometrika, 25,* 313-323.

Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. *Journal of Educational Psychology, 59,* 375-380.

Humphreys, L. G. (1986). Stability and instability of individual differences. In *FY86 Annual Report Supplement* (ARI Research Note 813704). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Humphreys, L. G., Park, R. D., & Parsons, C. K. (1979). Application of a simplex process model to six years of cognitive development in four demographic groups. *Applied Psychological Measurement, 3,* 51-64.

Humphreys, L. G., & Taber, T. (1973). Postdiction study of the GRE and eight semesters of college grades. *Journal of Educational Measurement, 10,* 179-184.

Husen, T. (1951). The influence of schooling upon IQ. *Theoria, 17,* 61- 88.

Jensen, A. R. (1981). *Straight talk about mental tests.* New York: The Free Press.

Kangas, J., & Bradway, K. (1971). Intelligence at middle-age: A thirty-eight-year follow-up. *Developmental Psychology, 5,* 333-337.

Kelly, E. L. (1955). Consistency of the adult personality. *The American Psychologist, 10,* 659-681.

Kuder, F., & Diamond, E. E. (1971). *Occupational Interest Survey: General manual* (Second Edition). Chicago: Science Research Associates, Inc.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95,* 179- 188.

Lin, P. C., & Humphreys, L. G. (1977). Predictions of academic performance in graduate and professional school. *Applied Psychological Measurement, 1,* 249-257.

Lorge, I. (1945). Schooling makes a difference. *Teachers College Record, 46,* 483-492.

Mayberry, P. W. (1988, April 26). *Marine Corps analysis of new predictors (CNA 88-0748.10).* Alexandria, VA: Center for Naval Analyses.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89,* 191-216.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25,* 707-726.

Moss, H. A., & Susman, E. J. (1980). Longitudinal study of personality development. In O. G. Brim, Jr. and J. Kagan (Eds.), *Constancy and change in human development* (pp. 530-595). Cambridge, MA: Harvard University Press.

Mumford, M. D., & Owens, W. A. (1982). Life history and vocational interests. *Journal of Vocational Behavior, 21,* 330-348.

Owens, W. A. (1966). Age and mental abilities: A second follow-up. *Journal of Educational Psychology, 57,* 311-325.

Peterson, N. G. (Ed.). (1987). *Development and field test of the trial battery for Project A* (Report No. TR-739). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Plant, W. T., & Telford, C. W. (1966). Changes in personality for groups completing different amounts of college over two years. *Genetic Psychology Monographs, 74,* 3-36.

Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100,* 67-77.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24,* 419-434.

Shaffer, G. S., Saunders, V., & Owens, W. A. (1986). Additional evidence for the accuracy of biographical data: Long-term retest and observer ratings. *Personnel Psychology, 39,* 791-809.

Siegler, I. C., George, L. K., & Okun, M. A. (1979). A cross-sequential analysis of adult personality. *Developmental Psychology, 15,* 350-351.

Strong, E. K., (1943). *Vocational interests of men and women.* Stanford, CA: Stanford University Press.

Sussmann, M., & Robertson, D. U. (1986). The validity of validity: An analysis of validation study designs. *Journal of Applied Psychology, 71,* 461-468.

Tuddenham, R. D., Blumenkrantz, J., & Wilkin, W. R. (1968). Age changes on AGCT: A longitudinal study of average adults. *Journal of Clinical and Consulting Psychology, 32,* 659-663.

U.S. Department of Labor, Manpower Administration. (1970). *Manuals for the USTES General Aptitude Test Battery, Section III: Development.* Washington, DC: Author.

Vernon, P. E. (1957). Intelligence and intellectual stimulation during adolescence. *Indian Psychological Bulletin, 2,* 1-6.

Vernon, P. E. (1979). *Intelligence: Heredity and environment.* San Francisco: W. H. Freeman and Company.

Wahlstrom, M., & Boersman, F. J. (1968). The influence of test-wiseness upon achievement. *Educational and Psychological Measurement, 28,* 413- 420.

Wheeler, L. R. (1942). A comparative study of the intelligence of East Tennessee mountain children. *Journal of Educational Psychology, 33,* 321-334.

Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year* (College Board Report No. 83-2). New York: College Entrance Examination Board.

Wolfe, J. H. (1988). Future tests: Design for validation in ten Navy schools. *Proceedings of the 30th Annual Conference of the Military Testing Association,* Arlington, *VA.*

Yates, A. J., James, W. S., Dempster, J. J. B., Wiseman, S., & Vernon, P. E. (1953). Symposium on the effects of coaching and practice in intelligence tests. *British Journal of Educational Psychology, 23,* 147-162 (also *24,* 1-8, 57-63).

Distribution List

Headquarters, U.S. Military Entrance Processing Command (MEPCT-P)
Office of the Assistant Secretary of Defense (F&P) (2)
Director, Research and Development Department of Defense Coordinator
Deputy Chief of Naval Operation (M&P), (N091), (N1), (N1B)
Director, Recruiting and Retention Program Division (PERS-01JJ), (PERS-23)
Defense Technical Information Center (DTIC) (4)